CATEGORISATION FOR SMALL–MEDIUM INFORMATION SYSTEMS—AN EXPLORATION

JAMES SINCLAIR

A thesis submitted for the degree of Doctor of Philosophy of the Australian National University

Department of Engineering
Faculty of Engineering and Information Technology
The Australian National University

July 2007

DECLARATION

This PhD research has been conducted under the supervision of Professor Michael Cardew-Hall, Dr Eric McCreath and Professor David Hawking.

I certify that this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university, and that, to the best of my knowledge, it does not contain any material previously published or written by another person except where due reference is made in the text. The work in this thesis is my own.

The Australian National University
Canberra, July 2007

James Sinclair

So, whether you eat or drink, or whatever you do, do all to the glory of God.
— 1 Corinthians 10:31

To the Glory of God, and for my wife, Summer.

ABSTRACT

This thesis is an exploratory study, investigating the causes and mechanisms of categorisation problems in information systems. Looking at both cognitive functions in the brain and the context of information systems shows that categorisation is far from simple. Individuals vary so greatly as to make the design of a perfect categorisation scheme impossible. At the same time however, category structures in the mind are not arbitrary or random, and there are many commonalities between people. Hence a good categorisation scheme will find a balance between accommodating individual differences and encouraging conformity.

The investigation process revealed a gap in the literature concerning assumptions about how to solve the problem. Most proposed solutions assume that an expert administrator is available to maintain category structures, that the items to be categorised will be primarily textual, and that the dataset will be very large. The reality is however, that these assumptions do not always hold true. This thesis proposes that by using tag clouds and clustering techniques, folksonomies can be adapted to suit smaller information systems where no dedicated administrator is available to maintain the category scheme. This was demonstrated through a number of experiments evaluating these approaches.

PUBLICATIONS

The following publications have been produced as a result of this work:

Sinclair, J. and Cardew-Hall, M. [2007], 'The folksonomy tag cloud: When is it useful?', *Journal of Information Science* (In press).

Sinclair, J. and Rossiter, M. [2005], Graded categories for knowledge management systems, *in* N. Pujawan, ed., '1st International Conference on Operations and Supply Chain Management, Bali, 2005'.

Don't be deceived, my dear brothers.
Every good and perfect gift is from above,
coming down from the Father of the heavenly lights,
who does not change like shifting shadows.

— James 1:16-17

ACKNOWLEDGMENTS

First and foremost, I would like to thank my wife Summer, who has put up with a great deal through the production of this document. I would also like to thank my supervisors Michael Cardew-Hall, Eric McCreath and David Hawking for all you help and feedback. Thanks also to Victor Pantano and Margaret Rossiter for encouraging me and providing me with good advice when I needed it. Bob Field very kindly offered to read and edit my entire manuscript and offered excellent suggestions for improvement, for which I am most grateful.

CONTENTS

LIST OF TABLES

ACRONYMS

ANU Australian National University

CAD Computer Aided Design

CBR	Case-Based Reasoning
CS	category scatter
CIO	Chief Information Officer
CSCW	Computer Supported Cooperative Work
CRG	Classification Research Group
CWA	Cognitive Work Analysis
DoM	degree of membership
DVD	Digital Versatile Disc
GI	Graded Interface
FE	Finite Element
FAST	Faceted Analytico-Synthetic Theory
FMI	Future Model Improvement
GOMS	Goals, Operators, Methods and Selections
HCI	Human–Computer Interaction
HTML	HyperText Markup Language
ICD	International Classification of Diseases
ICS	intra-cluster similarity
IR	Information Retrieval
ISBN	International Standard Book Number
ISSN	International Standard Serial Number
KM	Knowledge Management
k-NN	k nearest neighbours
LIS	Library and Information Science
LCSH	Library of Congress Subject Headings
NGI	Non-Graded Interface
PDF	Portable Document Format
SME	Small–Medium Sized Enterprise
SVM	support vector machine
TF-IDF	Term Frequency—Inverse Document Frequency
URL	Uniform Resource Locator
WWW	World Wide Web
XML	eXtensible Markup Language

1

INTRODUCTION

Categorisation is fundamental to how we function as human beings, and pervades every aspect of our lives. It profoundly shapes our cognitive processes, the world we live in, and our social interactions. Most of the time, we categorise unconsciously and effortlessly yet, for some reason, information is particularly difficult to categorise well.

Categorisation forms the basis of almost everything we do. 'There is nothing more basic than categorization to our thought, perception, and speech' [70]. Categorisation allows us to interact with the world and with other people, and forms the basis of our thought processes. As Estes [35] writes:

> I will not dwell at length on the importance of classification and categorization in the cognitive domain; how difficult our lives would be if we did not have classification skills; how many kinds of activities have classification at their heart. Suffice to say that classification is basic to all our intellectual activities.

Categorisation is not only the basis for our cognitive functions. It also pervades the world we have created for ourselves. Enter a kitchen and you will usually find cutlery kept together, yet spoons will be separated from knives, and knives from forks. Food is kept away from poisons, yet grouped with other foodstuffs. Cities are divided into industrial zones, business districts and residential areas. The goods on the supermarket shelves are grouped into categories, as are DVDs in the rental store. Our salaries as employees are determined by our job classification, and the category of vehicle we drive determines how much we pay at a toll gate. Categories surround us everywhere we go.

Not only are we surrounded by categories, but we also spend a great deal of time categorising. 'We all spend large parts of our days doing classification work, often tacitly, and we make up and use a range of ad hoc classifications to do so' [16]. In the laundry, we separate dirty washing from clean, and light-but-dirty washing from dark-but-dirty washing; in the garden, weeds are separated from other plants; and when we travel, we create all kinds of categories such as

things to take in my hand luggage or *places I wish to avoid*. Our daily activities are full of categorisations.

Categorisation also shapes our social interactions and even our own self-perception. We use categories to identify who we are, and what we should do. McGarty [89] illustrates this using the example of a football match:

> Imagine you are going to watch a football match at a stadium between a team you support and a traditional rival. In order to understand the game you would, at the very least, need to categorize the players as belonging to different teams. To avoid being arrested you would need to categorize yourself as a spectator and not as a player. These categorizations are relatively obvious and may require little if any conscious thought on your part. More interestingly, however, you may come to categorize yourself as a supporter of one of the teams.
>
> If you are like most supporters, as incidents occur on the field you will come to classify decisions by the referee or umpire as fair or unfair (and hence to be met with silence or derision) and segments of play as worthy of comment, applause or silence. You may well notice that many of your classifications seem to be shared by other people who support the same team as you.
>
> However, you could also hardly fail to notice that the classifications that you share with other supporters of your team seem to be keenly contested by the opposition supporters. They seem to classify fair decisions as worthy of derision and often greet examples of the most scintillating play with stony silence. However, rather than being puzzled by this disagreement we actually expect this perverse behaviour from the opposition.

Most of the time, we categorise unconsciously and effortlessly. It is just *what we do*. We are so good at categorising that we don't even think about it. Yet when it comes to organising information, categorisation seems to be more difficult. Things don't quite seem to fit. We often can't find the one small piece of information we need—it's not where we expect it to be.

One study [80], reports on a series of interviews examining how people organise information in their desks and offices. In the analysis, the author grouped participants into those with 'messy' offices,

and those with 'neat' offices. Not surprisingly, the report found that people with messy offices had more difficulty finding information than people with neat offices. However, keeping an office organised required a great deal of effort. '[O]ne of the people with a neat office said that he had spent over eight hours organizing the filing system in his desk drawer but that doing so had not been as valuable as he expected.' The eight hours of effort hardly seemed worth the payoff. If we are so good at categorisation, why is it that organising *information* is so difficult?

The sheer volume of information people deal with every day further exacerbates the problem, particularly in an enterprise context. 'We live in an information society in which more people must manage more information, which in turn requires more technological support, which both demands and creates more information' [83]. We are bombarded with information, but not all of it is the information that we need, as Edmunds and Morris [34] write:

> 'Although there is an abundance of information, it is often difficult to obtain useful and relevant information among the vast volumes of information'.

We have an increasing need for information, and vast amounts of information are available, yet it is not the information we need.

Why then, is information so difficult to categorise well? One reason is that information is constantly changing. New information keeps coming in, and the new information changes the meaning of information already there; some information becomes more important while other information becomes obsolete. Our minds are well equipped to deal with this kind of change, constantly updating and modifying category structures as needed. Information systems however, are generally not so well equipped to deal with change. Unless someone (or a team of people) regularly updates and maintains the category structure then problems will occur: When today's information must be squeezed into yesterday's category structures it does not always fit.

Another reason for categorisation difficulties is that people's conceptual structures are very different from each other. Our brains adapt and optimise themselves to suit our specific circumstances, such that no two people see the world entirely alike. Information systems on the other hand, are generally used by many different people, each with their own unique view of the world. And not only do people's conceptual structures differ, but also the language they use to de-

scribe concepts. This makes categorisation in any information system difficult because different people will use different words to describe the same thing.

Of course, the categorisation problem is not a new one. The entire body of literature in Library and Information Science (LIS) exists to study this problem of how best to categorise and organise information. Quite understandably however, much of the LIS literature assumes the presence of a librarian or information scientist. And indeed, many large organisations employ people in roles such as librarian or Chief Information Officer (CIO) specifically to deal with organising information. If information is so important, and categorising information is so difficult, it makes sense to employ an expert in the area.

Not everyone can afford the luxury of a dedicated librarian however, particularly in Small–Medium Sized Enterprises (SMEs). Employing an expert is expensive, and organisations often perceive information management as supportive of, rather than central to, the core business of the organisation (see Chapter 2). Even within large organisations, the organisation-wide information systems may not be well suited to the needs of individual departments or sections. Furthermore, these individual departments or sections may not have the resources to employ their own librarian.

Where no dedicated librarian is available, full-text searching is often used to help people find the information they need. In the last decade, Information Retrieval (IR) tools and techniques have become increasingly ubiquitous. Search engines like Google have been extremely successful and made searching a familiar activity in many people's daily lives. Yet the IR literature also tends to make assumptions that do not always hold in smaller organisations. Firstly, it is assumed that the data is primarily textual. Secondly, the literature also tends to assume very large amounts of data. For the World Wide Web (WWW), these assumptions have held very well—the WWW does indeed contain vast amounts of textual data. In the context of work groups and SMEs however, the amount of data to organise may not be massive, but it is large enough to require management. Furthermore, as seen in Chapter 2, the data collected is not always primarily textual.

In order to address these gaps, we formulate the following problem definition. That is, the aim of this thesis is to find categorisation methods that:

A. reflect user vocabulary;

B. are able to adapt to change;

C. do not require expert human intervention;

D. are suitable for multimedia data; and

E. work effectively with small numbers of records.

1.1 THESIS OVERVIEW

This thesis is divided into two parts. Part I explores the categorisation problem in detail, looking at reasons why the categorisation problems occur, beginning with an enterprise case study. This gives the background and reasoning for our problem definition. Part II then presents *folksonomies* as a potential solution to the problem. In particular, Part II explores ways of utilising folksonomy data to best advantage and shows how these techniques might be applied in a small–medium sized information system.

Part I: The Categorisation Problem

CHAPTER 2 uses a case study to explore the categorisation problem in an enterprise context. The issues described give motivation for the thesis and further clarify the problem addressed.

CHAPTER 3 presents reasons for the categorisation problem arising from the way categorisation occurs in the brain. The way categories are structured in the brain is significantly different from the way we normally organise information systems. This leads to a number of issues and implications for information system design.

CHAPTER 4 reports on a study investigating the implications of Chapter 3 for user interface design. Based on the findings of Chapter 3, Chapter 4 presents the results of an experiment comparing two user interfaces: one which allowed for graded category membership, and one which allowed only binary membership. The study found that a change in user interface can result in greater categorisation consensus, but this is offset by an increase in time required to categorise.

CHAPTER 5 presents reasons for the categorisation problem arising from the context in which information is categorised. Cognitive is-

sues do not account for all the effects of the categorisation problem. Many issues arise from the context in which categorisation occurs. In fact, contextual issues arguably result in much more significant effects than cognitive ones. After reviewing these issues, Chapter 5 then formulates a specific problem definition.

CHAPTER 6 gives an overview of attempts to address the categorisation problem in the literature. I then propose folksonomies as a potential solution to the categorisation problem.

Part II: Folksonomies

CHAPTER 7 reports on a study investigating user interfaces for folksonomies. In particular, it explores the use of *tag clouds* as a means of navigating folksonomy-based data.

CHAPTER 8 presents a study on clustering folksonomies to address some of their disadvantages as an information organisation method. Many authors and practitioners have implemented various means for clustering folksonomies, yet very little justification is given for the choice of one technique over another. Chapter 8 presents a comparative study of clustering techniques applied to folksonomies, showing that a technique not previously applied to folksonomies produces superior clustering results.

CHAPTER 9 reports on the design of a folksonomy-based system, showing how the techniques explored in Chapters 7 and 8 can be applied in a real life setting.

CHAPTER 10 provides a summary of the research and key findings. It also presents possible directions for future research in this area.

1.2 ORIGINAL CONTRIBUTIONS

In exploring the categorisation problem, this thesis makes a number of original contributions:

1. This thesis provides a cross-disciplinary investigation of the causes of categorisation problems. Beginning with the case study in Chapter 2, the thesis explores causes of categorisation problems, ranging from cognitive issues arising from how our brains

operate, to issues arising from the context in which categorisation occurs. The analysis incorporates literature from cognitive science, psychology, LIS, knowledge management, Human–Computer Interaction (HCI), and computer science. Although other analyses [e.g. 16, 62, 70] can legitimately claim to be cross-disciplinary, they tend to focus either on cognition or on context. Through the case study in Chapter 2 this thesis ties these two aspects together, showing that both play a part in making categorisation difficult.

2. The study in Chatper 4 reports on a study comparing a graded interface (slider-bars) with a non-graded interface (radio buttons) for categorisation. The results showed that using the graded interface increased accuracy by around 3%; however, people using the graded interface took, on average, 8.8 seconds longer to categorise. The results also showed that individual people not particularly good at predicting the categorisations of their peers, with pair-wise correlations ranging from 0.35 to 0.47. In spite of this, the data suggests that the collective contributions of users can be aggregated to counteract individual disagreement. Hence in systems with multiple users, collaborative categorisation may be a viable alternative to employing an expert administrator.

3. After presenting folksonomies as a potential solution to the categorisation problem, this thesis then addresses gaps in the folksonomy literature. Chapter 7 addresses the issue of tag clouds. As with many issues surrounding folksonomies, much of the debate over tag clouds has occurred in the blogosphere, which 'presents a difficult challenge to researchers in terms of properly evaluating and acknowledging contributions that have not been externally vetted' [18]. Chapter 7 presents empirical evidence on the usefulness of tag clouds as a user-interface element, finding that the tag cloud does indeed provide value to people seeking information from a folksonomy data set. Specifically, the tag cloud has a number of positive attributes: A) It is particularly useful for browsing or non-specific information discovery, B) the tag cloud provides a visual summary of the contents of the database, and C) it appears that scanning the tag cloud requires less cognitive effort than formulating specific query terms.

   The tag cloud is not without its limitations however. A corollary

of the tag cloud's suitability for general browsing is its unsuitability for seeking specific information. This means that the tag cloud is not sufficient as the sole means of navigating a folksonomy dataset.

4. Chapter 8 addresses another gap in the folksonomy literature regarding the use of clustering techniques with folksonomy data. Many authors and practitioners have implemented various means of clustering folksonomies. However, each approach is different and utilises different clustering techniques. Very little justification is given why one technique is chosen over another. Chapter 8 presents a comparative study of different clustering techniques applied to folksonomies. The study compared four clustering algorithms against two folksonomy datasets. Of the four algorithms tested, the ROCK algorithm performed best, but was also the most algorithmically complex. All four algorithms showed significant improvement over the random baseline, indicating that folksonomies do capture semantically valuable data.

5. Chapter 8 also introduces a novel approach for comparing the results of a clustering algorithm with a human-generated categorisation scheme. In testing, I found that existing clustering quality measures gave maximum values at either one cluster or N clusters (where N is the number of items). This is undesirable, since having one or N clusters is essentially the same as not clustering at all. The approach described in Chapter 8 enables a non-mutually-exclusive category set to be compared with a mutually exclusive categorisation scheme, without giving maximum values at one or N clusters.

Part I

THE CATEGORISATION PROBLEM

2

CASE STUDY: CATEGORISATION IN A MANUFACTURING ENVIRONMENT

Is categorisation really that difficult? How do we even know there is a problem? Often, we have vague sense that we can't find what we want, or things aren't organised well, but what gives us this impression? What does the categorisation problem look like in an organisational context?

To investigate these questions, I conducted a study of a Knowledge Management (KM) system, called SIMPRESS, developed by the Australian National University (ANU) for an Australian automotive manufacturer. SIMPRESS is a tool that allows shop-floor operators in a sheet metal forming plant to record how they solve problems. Part of the data entry process involves categorising the problem. After examining these categorisations the results showed a number of problems symptomatic of poor categorisation. Large numbers of entries were categorised as *Other*, while many of the categories provided were not used at all. The majority of entries categorised as *Other* contained textual information showing that they clearly belonged in a defined category. In short, the category system in SIMPRESS was almost useless as a method for organising information and facilitating knowledge re-use.

Given these problems, a natural question to ask is why they occur. Analysis of the user interface and interviews with operators revealed that some causes of these problems have little to do with the system itself. Rather, the organisational context and culture plays a significant role in determining how the system is used. The categorisation problems are not simply the result of poor system design. But neither is the system design irrelevant. System design affects how people perceive the system, which also plays a role in determining how the system is used. Thus, any investigation into categorisation in information systems must examine both information systems and the people who use them.

2.1 BACKGROUND

Understanding the categorisation problems in SIMPRESS first requires some understanding of the manufacturing environment in which it

was implemented. SIMPRESS was developed for an Australian automotive manufacturer to support operations in a sheet metal stamping plant. So, before describing SIMPRESS, I briefly review the sheet metal forming process.

2.1.1 *Sheet Metal Forming*

Sheet metal forming is a process where a series of heavy presses (Figure 1) progressively stamp a flat metal sheet (called a blank) into a three dimensional shape. The molds used to form the shape in each press are called dies. Sheet metal parts form the skin panels and body shape of most consumer vehicles produced today.

Figure 1: A line of heavy presses used in sheet metal forming.

Before a sheet metal part is put into production, engineers must design both the part and the manufacturing process to make it. Experienced craftsmen then build the dies used to stamp the desired shape from each blank. Figure 2 illustrates this product life-cycle of for a sheet metal part. Listed below is a short description of each stage:

PART DESIGN is where the final shape of the metal part is designed. The part must fit the styling and specifications given by the product design team that develops the concept for the overall design of the vehicle.

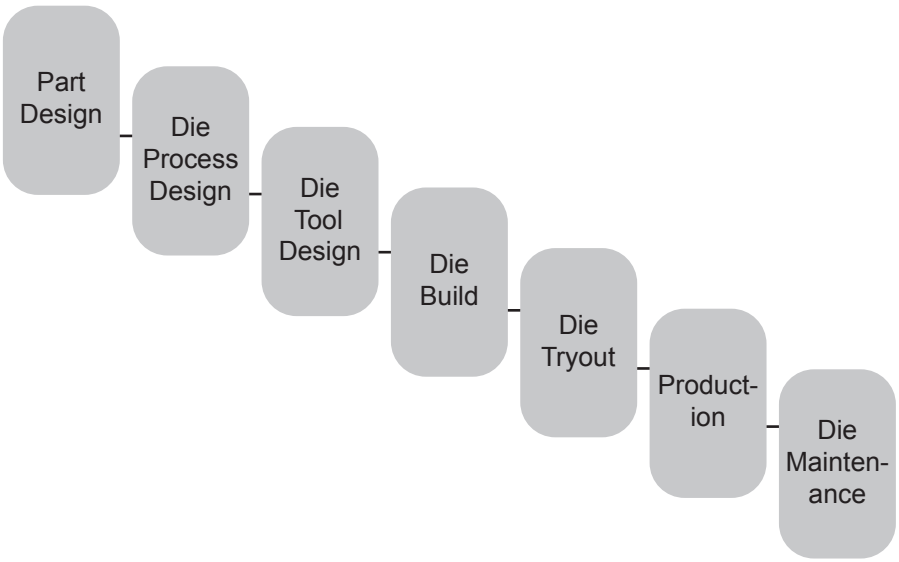


Figure 2: Activities involved in the sheet metal forming product life cycle (adapted from Smith et al. [123]).

DIE PROCESS DESIGN is where engineers determine how the part will be made. Usually the final shape of a part cannot be achieved in a single operation, but must be broken down into a number of stages. Engineers must determine the shape of the initial blank, the number of presses required and how parts will be moved between presses.

DIE TOOL DESIGN. Once the manufacturing process has been determined, the dies for each press operation must be designed.

DIE BUILD is where the die tooling is actually built. Usually this process begins by casting large steel blocks into an approximate shape. The die is then machined and polished to produce the exact shape desired.

DIE TRYOUT is where the dies are tested by attempting to produce a stamped metal part. Very rarely will a set of dies immediately produce satisfactory parts first time without some modification.

PRODUCTION is when parts are actually made. Even here, variability in the production process and wear on dies can result in faulty parts being produced, so die maintenance must be carried out.

DIE MAINTENANCE. As mentioned, wear on dies can lead to problems in production, so dies must be maintained to ensure quality.

Metal stamping is a process that relies heavily on the experience of skilled staff, particularly in the building and tryout of dies. In spite

of increased use of Finite Element (FE) analysis and other numerical methods, the tryout and production phases of the product life-cycle are often problematic. That is, problems with dies become apparent on the shop floor, in the tryout and production phases, not on the engineers' desks. These problems are resolved using a trial-and-error process, which can be both costly and time consuming to fix [23]. Hence, metal stamping is sometimes referred to as a 'black art', characterised by rules of thumb and the experience of sages, rather than formalised procedures and scientific method.

2.1.2 *The Knowledge Management System*

SIMPRESS differs from many other KM systems designed for sheet metal manufacture because it is designed for use by shop-floor operators, rather than design engineers. That is, SIMPRESS was designed to utilise the significant skills and experience of shop-floor operators at the tryout and production stages of the product life-cycle [123]. This is in addition to the expert knowledge of highly-trained engineers in the earlier design phases.

The aim of the system was to record problems that occur during tryout and production, along with a description of how they were solved. In this way, when similar problems are encountered later, operators can review how they were solved. These records are also fed back upstream to the design engineers, so that problems can be 'designed out' (Figure 3).

Developing a KM system for use by shop-floor operators presented unique challenges. Unlike a KM system in more traditional 'knowledge worker' environments (such as law, software design, R&D, etc.), workers in this environment tended to have low levels of computer literacy. In some cases the introduction of a new IT system was viewed with deep suspicion [103]. Thus, SIMPRESS needed to fill a number of requirements [124]:

- It needed to be simple and intuitive to use, even for people with low levels of computer literacy.
- The system needed to reflect the tasks and reasoning processes of the shop-floor operators. This included the language and procedures already existing in the manufacturing environment.
- Similarly, the system needed to support (or be supported by) the daily business practices of the organisation; in this case, the

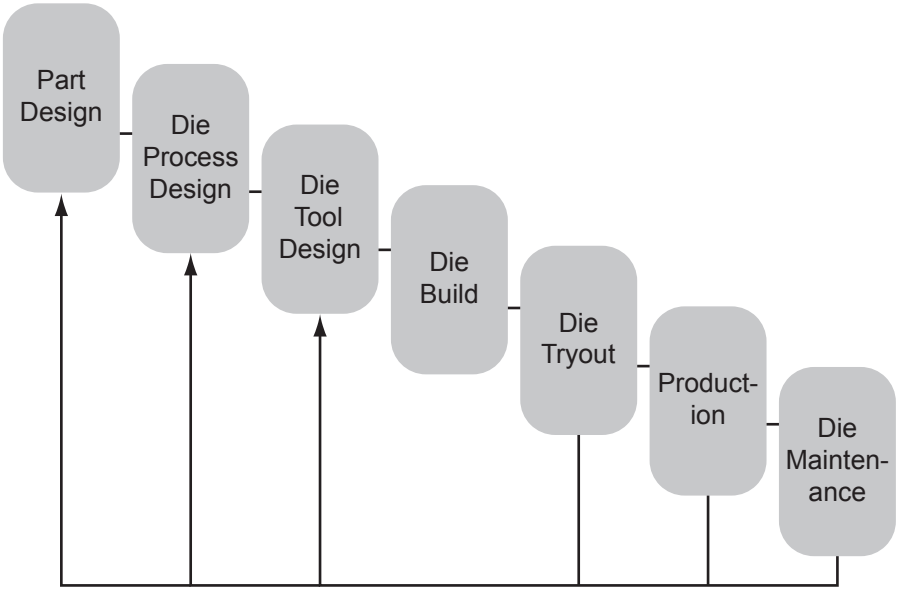


Figure 3: Knowledge feedback loop facilitated by SIMPRESS

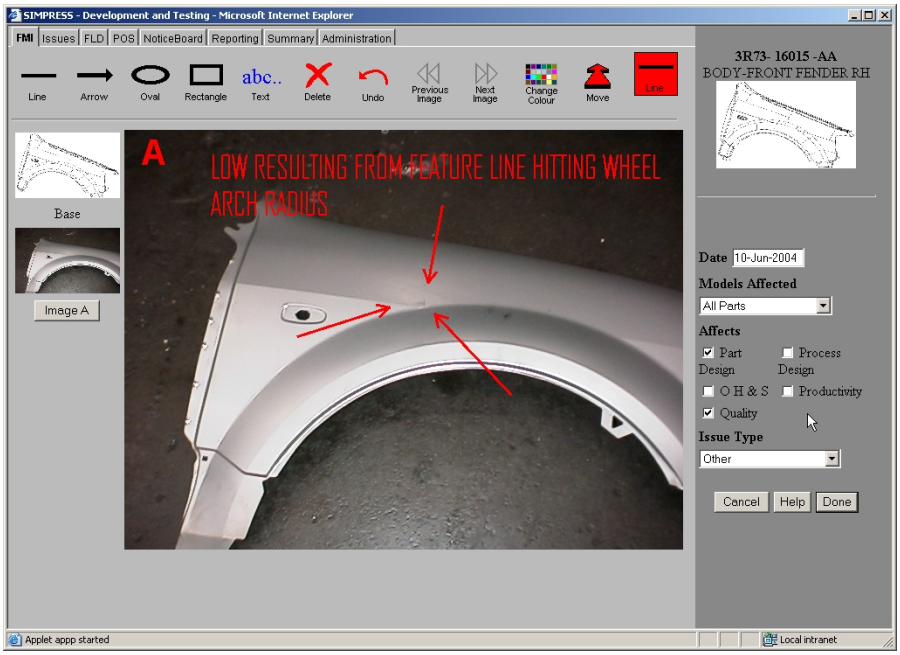
production of metal parts.

To meet the first of these three requirements, the developers of SIMPRESS put a great deal of effort into integrating images such as CAD drawings and digital photographs into the system. Instead of writing a complicated textual description of where and what problem was occurring, an operator could simply take a photo of the part, upload the image, and draw a red circle around the problem area. With this in place an operator can write a shorter, simpler textual description (such as 'splitting in this area due to...') without having to describe the exact location in words.

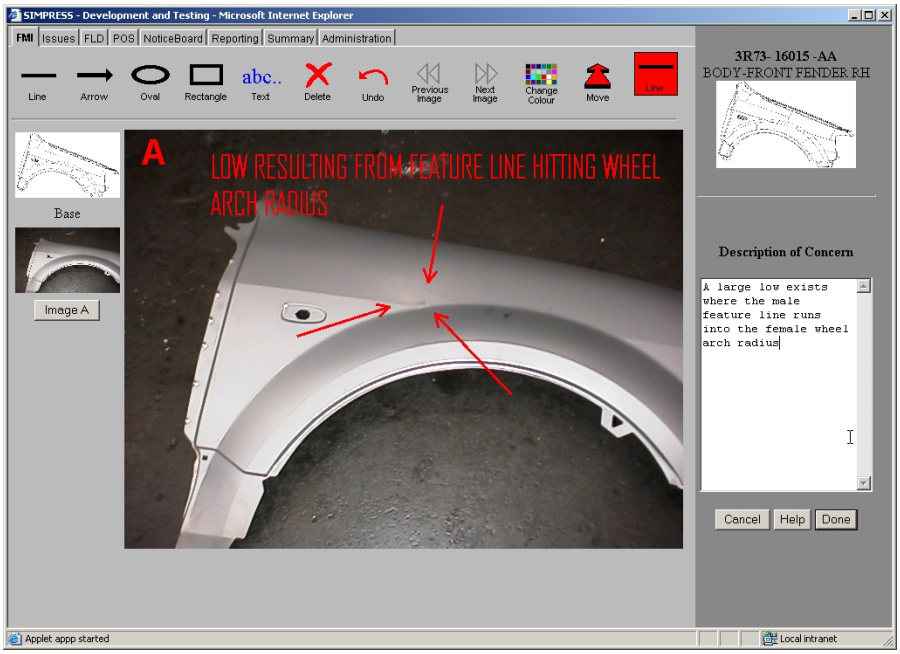
To support the daily business practices of the organisation, SIMPRESS introduced the Future Model Improvement (FMI) module. This module replaced an established paper-based procedure of reporting problems to design engineers. The procedure consists of a shop floor operator writing a description of a problem encountered in tryout or production, with a suggestion to engineers for preventing the issue in future designs. Responsibility for the FMI is then assigned to an engineer for review at certain stages of the product design process. In this way, SIMPRESS became integrated into existing plant procedures. More detail on the design and implementation of SIMPRESS can be found in Cardew-Hall et al. [23], Cardew-Hall et al. [24], Smith et al. [124] and Smith et al. [123].

2.1.3 *Categorisation in* SIMPRESS

Figure 4 shows two of the data screens for entering an FMI in SIMPRESS. On the first of these screens there is a field where operators can select a *concern type* from a drop-down menu. There are 53 concern types listed in the drop down menu (see Appendix A), to cover the wide range of issues that may occur in production and tryout. The concern types were carefully selected by developers to reflect the vocabulary of operators after a thorough domain analysis. In spite of this careful consideration, however, the developers expressed a concern that there were problems with the categorisation being performed in SIMPRESS. The aims of this study were to determine exactly what kinds of problems were occurring and explore why this was the case.



(a) Data entry screen showing *Concern Type* drop-down menu.



(b) Data entry screen showing textual description box

Figure 4: Data entry screens for loading an FMI in SIMPRESS

2.2 THE STUDY

In order to investigate the categorisation problems occurring in SIMPRESS, I used three sources of data:

1. Tabulation of categories used in all FMI entries in SIMPRESS from December 2001 to November 2004.
2. Printed copies of 201 full FMI entries entered between 15th October 2003 and 11th June 2004.
3. Semi-structured interviews with personnel who had entered the FMIs and other stakeholders (conducted between 31st January to 2nd February 2005).

The category tabulations were collected using simple database queries to measure the number of entries associated with each category. Counts were also taken of the number of entries made per month, and the number of entries made by each department within the organisation.

The hard-copy FMI entries were examined by reading through each entry, noting the personnel involved, concern type category, associated problem descriptions and any digital images. This was followed by a second reading of the reports, noting misclassifications, and any classification information included in the textual description.

The results from the category tabulation and analysis of hard-copy FMIs are presented in Section 2.3. These represent results that were obtained solely by analysis of data taken directly from SIMPRESS, and present symptoms of categorisation problems occurring in the system. To examine why these problems were occurring, I conducted interviews with people who use the system in a range of capacities.

In total, seven interviews were conducted with people who had entered the FMIs analysed previously in hard-copy. In some cases interviewing people who had entered FMIs was impossible, as some employees had left the organisation. Where this was the case, I discussed SIMPRESS with the manager responsible for the former employees. I also interviewed some of the design engineers who receive FMIs from people further downstream in the product life cycle.

Indicative questions from the interviews are listed in Table 1. Originally, I had planned to perform two rounds of interviews: one focussing on the SIMPRESS system in general, and one focussing on categorisation issues. However, after performing the interviews, it was clear that the operators would have resented the further imposition

on their time. While they were happy to talk with me for one interview, it was clear that most expected not to be bothered further. Hence the questions are more general than would have been ideal. The interviews did, however, reveal a number of issues affecting the use of SIMPRESS within the organisation. These issues shed light on many of the categorisation problems observed in the numerical analysis.

Table 1: Indicative questions for the semi-structured interviews.

1.	Can you tell me about your job in the organisation?
2.	Do you use SIMPRESS very often?
3.	What do you generally use SIMPRESS to do?
4.	When you make entries into SIMPRESS, are they generally things you come up with, or does someone else ask you to put them in?
5.	Do you ever need to go back and find entries in SIMPRESS? If so, how do you go about doing it?

Interviews were not tape-recorded as the noise in various sections of the manufacturing plant would have made this impossible. Instead, I recorded detailed notes and observations as soon as possible after the interviews were conducted; usually within 15–30 minutes of the interview. The interview notes were coded and analysed to identify key themes affecting categorisation in SIMPRESS. The results of this analysis are presented in Section 2.4.

2.3 RESULTS FROM SIMPRESS

The numerical counts, and analysis of hard-copy FMI entries revealed a number of key indicators showing that categorisation was not being performed well in the system. These included:

- Large numbers of entries categorised as *Other*; and
- Uneven category usage.

The sections following discuss each of these issues in turn, and then discusses potential sources for these problems arising from the system design.

2.3.1 *Use of the* Other *Category*

Table 2 shows the usage frequency for each concern type for *all* the FMIs in the system up to November 2004. At that time, a total of 449

entries had been made, of which 318 were categorised as *Other*—over 70% of entries.

Table 2: Usage frequency for each concern type.

CONCERN TYPE	ENTRIES	CONCERN TYPE	ENTRIES
Other	318	Splits	2
Design	49	Variation	2
Tolerance	15	Weakness	2
Burrs	8	Thinning	1
Measurement	5	Timing	1
Misalignment	4	Trim / Flange	1
CAD	4	Trim Edge	1
Fouling	4	Slug Build-up	1
Weld Integrity	4	Streamers	1
Springback	3	Damage	1
Bedding	3	Gripper	1
Location Pins	3	Lifter	1
Mislocation	2	Light-On	1
Safety	2	Scores	1
Buckles	2	Lows	1
Scrap Build-up	2	Flange Clearance	1
Tear Outs	2		

In the case of the hard-copy FMIs, the percentage of entries classed as *Other* was not quite so high. Of the 201 entries examined in detail, 55% of the entries were placed in the *Other* category. However, this is still over half the entries analysed. Only 89 out of the 201 entries were classified with a specific concern type.

To investigate this further, the entries classified *Other* were broken down into groups as shown in Table 3. *Classifiable from text* indicates the number of entries that could easily be placed in the correct category, without requiring expert domain knowledge. Usually this occurred when the concern type was written in the textual description. *Expert* indicates entries where expert domain knowledge would be required to determine the correct categorisation. *Correctly other* gives the number of entries that clearly did not fit into any of the available concern types. Finally, *Should be deleted* indicates entries that were left over from training sessions or entered mistakenly and should be removed from the system.

The results here indicate that a majority of the entries labelled *other* should rightly be placed in another category. Since the concern type

Table 3: FMIs labelled *other*.

Classifiable from text	64	(57.1%)
Expert	29	(25.9%)
Correctly Other	16	(14.3%)
Should be deleted	3	(2.7%)
Total Other	112	(100.0%)

was clear in the textual description, this raises the question 'why were these entries labelled incorrectly?' This question is examined further in Section 2.3.3.

Over-use of the *Other* category was not the only categorisation problem observed, however. Even when the *Other* entries are ignored there is an uneven distribution of category usage.

2.3.2 *Uneven Category Usage*

Another indicator of categorisation problems is that of under-used categories. Of the 53 available categories (listed in Appendix A), nineteen had not been used at all in the entire three year period. Thirteen of the concern types were only used once. This means that only 21 of the available categories (less than half) were used more than once.

In and of themselves, under-used categories are not necessarily a problem. They may simply mean that none of those concern types has occurred yet. However, without any entries, unused concern types add no value to the system. They are only of potential value if and when a problem of that type occurs.

At the same time, some categories appear to be over-represented. Looking again at Table 2, we can see that the top three concern types have considerably more entries than any of the others. The *Other* category we have already dealt with, yet there appears to be a tendency for general categories to be used more frequently than specific categories. For example, the *Design* category has more than twice the entries of the next, more specific, category *Tolerance*. And the *Tolerance* category is considerably larger than the next, *Burrs*, category.

Again, this is not necessarily a problem, in and of itself. It may simply reflect the types of problems being entered into the system. However, when it comes to finding solutions to problems, one suspects that the *design* category is too general to be really useful. Given the overall philosophy of the system described in Section 2.1.2, it is difficult to imagine a shop-floor operator looking for solutions to a

design problem. From the perspective of a design engineer too, the category *design* would appear to add little value.

2.3.3 *System Design Issues*

Why do these categorisation issues occur? Before examining data collected from interviews, we first examine potential sources of problems arising from the design of SIMPRESS itself.

One factor that may contribute to the high numbers of *Other* entries is that *Other* is the default option when a new FMI entry is made. If the person entering the FMI ignores the concern type for some reason, then the entry becomes *Other* by default. This is likely to affect the number of *Other* entries, but does not account for why the person entering the FMI would ignore the concern type.

Another possible factor is that the system only allows one concern type to be associated with an FMI. For example, a number of entries contained the text '*CAD/Design issue*'. Both *CAD* and *Design* are valid concern types. Should the operator pick one at random, or label the entry *Other*? Again, while this may be a contributing factor, not all the entries labelled *Other* would fit into multiple categories. Many of the entries fit well into a single category, yet were still labelled *Other*.

Yet another factor may be the sheer number of categories listed. To illustrate, imagine an operator is entering a new FMI under the category *Variation*. To choose the *Variation* category from the drop-down menu, the operator has to scroll through a list of 53 entries to find the single category that he wants. It may be easier simply to choose a vague, general category like *Other* or *Design*, rather than waste time searching for a suitable category that may or may not exist.

While these design issues may be contributing factors to categorisation problems in SIMPRESS, they do not appear to fully account for all effects observed. To investigate the causes further, we examine the data from interviews with people using SIMPRESS.

2.4 RESULTS FROM INTERVIEWS

Analysis of the FMI reports showed that there were indeed categorisation problems in SIMPRESS, but this analysis does not reveal the whole story. Why do these problems occur? The interviews revealed three common themes emerging from the broader organisational environ-

ment:

1. Delegation of data entry;
2. Inadequate training and low computer literacy; and
3. Misunderstanding of purpose.

Delegation of Data Entry

The shop-floor operators and design engineers are busy people. Corporate 'right-sizing' and dealing with day-to-day issues means that these employees have a lot to do. This busyness forces employees to prioritise their activities, focussing on what they perceive as their core task. In this case, this usually means producing the required quota of stamped-metal parts. Activities which support (but do not directly contribute to) the core task are given a low priority. This issue was also identified by Pantano et al. [103] in a related study on organisational culture and technology diffusion.

From the interviews, it is clear that SIMPRESS is seen as a support tool, rather than an essential part of daily operations. Hence, entering data into SIMPRESS is a low priority. In addition to this, data entry work is often perceived as boring and mundane [16]. Operators and engineers may be aware of the importance of recording problems and solutions, but they do not wish to do it themselves. Even if they did, they do not have time to do so anyway.

This leads to many entries in SIMPRESS being made by proxy—the people with the relevant knowledge delegate data entry to someone else who does not have relevant expert knowledge. Four of the seven people interviewed mentioned this as a problem. Under organisational policy, FMIs cannot be raised by one person on behalf of another—the person who enters the FMI into the system has their own name associated with it. This creates a problem when someone wishes to ask questions about an FMI. The person who entered the data has their name listed in the system, but they know very little about the issue that they entered.

One interviewee said 'I normally enter FMIs on behalf of engineers who were too busy to raise them for themselves.' She said that this was not ideal because when she entered the FMI her name goes against it as the Initiator, even though she is not the one who has the knowledge about the issue.*

* Interview notes `2005-01-31#2`

In one instance, a pair of 'co-op' students were given a large pile of paper forms and told to enter them into SIMPRESS. They entered around 200 entries one month (see Figure 5) without really understanding any of the data they were entering. When the students left, the FMIs raised under their names had to be reassigned to another employee who had not made the entries in the system.

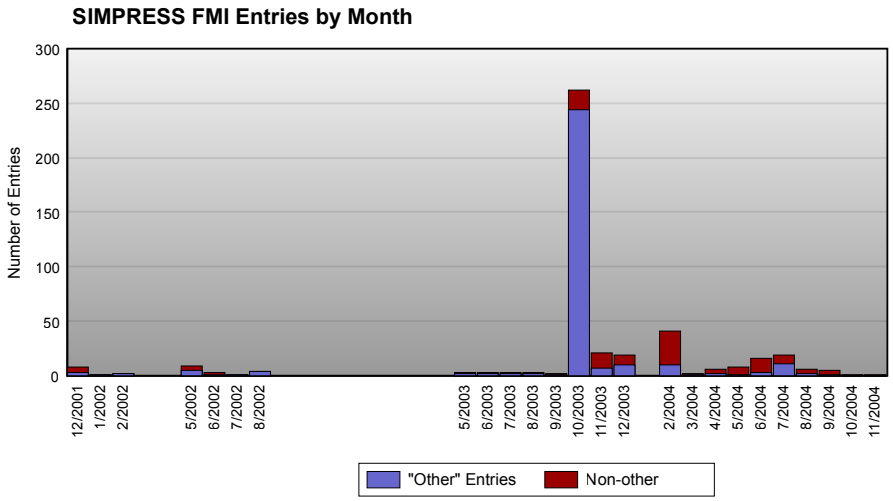


Figure 5: Number of FMIs raised by month

Given this situation where many entries are being made by proxy, people entering data often do not have adequate knowledge to categorise the entries properly.

Inadequate Training and Low Computer Literacy

Another factor contributing to poor categorisation in SIMPRESS is inadequate training given to some operators, and low levels of computer literacy. In one of the interviews an operator who entered FMIs on behalf of others indicated that 'some of the engineers weren't used to these kinds of things, like check-boxes and tables'.† Hence, she was delegated the task of data entry instead of those with the relevant knowledge. In a related study, Pantano et al. [103] also found that low levels of computer literacy hampered effective use of the KM system.

It was also made clear that the 'co-op' students mentioned above were not trained properly, and focussed on entering the data as quickly as possible. This meant that the concern type drop-down menu was largely ignored (indicated large percentage of *Other* entries in Figure 5), and a number of issues were entered multiple times.

† Interview notes `2005-01-31#2`

Misunderstanding of Purpose

As described by Smith et al. [123] the mindset in the plant is largely part-focussed. Part numbers are part of the common vocabulary of employees. To reflect this mindset, the system design is centred around part numbers. Most FMIs are logged against a specific part, and design reviews are carried out for individual parts, rather than to address different types of concerns.

In addition to this, FMIs are not viewed as a tool for knowledge re-use, but simply as a computerised version of a paper-based process. An FMI is a suggestion for preventing a problem occurring in future models of a car. Thus, it includes a problem description and suggested improvement entered after the problem has been encountered and solved. It is not a record of the problem solving process itself. Operators and engineers did not really consider the potential for knowledge re-use beyond formal design reviews for specific parts.

Thus, from the shop-floor operator's point of view, they have no reason to ever retrieve an FMI from the system. They enter the suggested improvement and the design engineers implement it. This was reflected in an interview with John,‡ a manager in stamping operations:

> When I asked if [John] ever needed to go back and look at FMIs, [John] said 'No.' To clarify, I asked if they just entered the FMI and then never saw it again. [John] said that the system sends him an email when it had been looked at by somebody and actioned. This seemed to me to mean that the initial raising of the FMI was the limit of his direct interaction with SIMPRESS.§

For the design engineers, the only time they would retrieve an FMI is when the design of a part is reviewed before a new model car is designed. Thus, there is no reason for them to search for any FMIs other than those associated with the part of interest. This was also reflected in one of the interviews:

> I asked if [Paul] would ever search by concern type and he said he might. While he was explaining it though, he seemed to change his mind and said that normally if they were reviewing acpFMI they would usually be interested

‡ Names have been changed.
§ Interview notes `2005-01-31#3`

> in a particular part and so would search on the part number. He said that in the future, members of different functional groups would be given a list of parts for their particular part of the car. They would have to go into SIMPRESS and look at any issues of FMIs loaded for each of those particular parts. He said that if they were doing that, they weren't likely to do a search for "splits", for instance, and then search through the list for any instances of the part, then repeat the process for each concern type.¶

Because the SIMPRESS FMI module is not viewed as a tool for knowledge re-use, the *concern type* categorisation is largely irrelevant to the people using the system. If the concern type is irrelevant, then there is no real incentive to make sure that the entries are categorised correctly.

2.5 DISCUSSION

If the concern type is irrelevant to the employees using the system, then are the categorisation problems also irrelevant? If no-one ever uses the *concern type* categorisation, then perhaps it should be removed from the system, since it adds an unnecessary task to the data entry process. The issue depends entirely on what the purpose of SIMPRESS actually is. If the FMI module is simply a more efficient version of a paper-based system, then yes, the *concern type* categorisation is irrelevant and should be removed. On the other hand, if the purpose of SIMPRESS and the FMI module is to encourage re-use of knowledge and assist in problem solving (as suggested by Smith et al. [123]), then some mechanism is needed to identify similar problems. Ostensibly, this is why the *concern type* exists.

If we assume that the *concern type* is important, then where does the blame for the poor categorisation lie? Is the user interface counter-intuitive and hard to use? Is it the prevalence of entry by proxy? Is it a result of inadequate training given to operators? Is it a failure to communicate the purpose of the system? The picture painted by our analysis and interviews suggests that the problem is not so simple as to have a single cause. The situation is complex, like most problems encountered when implementing information systems.

The solution to categorisation problems in information systems is not so simple as creating a better user interface. Issues such as organ-

¶ Interview notes `2005-02-02#1`

isational culture, training, and user perceptions of the system also have a significant impact on the quality of categorisation. And yet, the user interface is not unimportant either. The ability to enter data quickly and intuitively is critical when people are pressured for time. The user interface also affects people's perceptions of the system, which in turn affects their motivations in using it, and again impacts on categorisation quality.

Hence, any investigation of the categorisation problem should be multi-disciplinary. To get a full picture of the issue, one needs to understand how people interact with information systems, along with the cognitive processes behind categorisation. But even with a good understanding of these factors, categorisation problems cannot be understood in abstract from contextual issues such as those uncovered in our interviews.

2.6 CONCLUSION

As a case study, the SIMPRESS KM system illustrates two symptoms indicative of categorisation problems:

- Large numbers of entries categorised as *other*; and
- Uneven category usage.

Analysis of the user interface revealed that some problems may be the result of poor interface design. However, the interviews with people using the system revealed that there were a number of other significant factors such as:

- Delegation of data entry;
- Inadequate training and low computer literacy; and
- Misunderstanding of purpose.

These issues arise largely from the organisational context and culture. However, this is not to say that factors such as the system design and user interface are unimportant. It is clear that the categorisation problem is not the result of a single issue. There are multiple factors to be considered and any thorough analysis of categorisation in information systems will need to include several different disciplines of study.

3

COGNITIVE REASONS FOR CATEGORISATION PROBLEMS

To understand categorisation problems in information systems it is important to first understand categorisation at the cognitive level. This is an important step in understanding why categorising information is so difficult. This chapter reviews the work of cognitive scientists in understanding the structure of categories in the mind, and describes how this can cause problems when it comes to categorising information. I then outline some implications for the design of information systems.

A review of the literature revealed four main issues that lead to categorisation problems:

1. The way we commonly think about categorisation is somewhat different from the nature of categories in the mind. Section 3.1 outlines some of these commonly held views of categorisation. Section 3.2 then contrasts this with the work of cognitive scientists in understanding category structure in the mind.

2. The way we create structures for categorisation in information systems are also somewhat different from category structures in the mind. In particular, categorisation is often confused with *classification*, and hierarchical classification schemes are sometimes seen as the only right way to organise information. Section 3.3 describes the important differences between classification and categorisation.

3. Category structures in the mind are dynamic, continually updated and re-organised to accommodate new knowledge and understanding. While this happens somewhat automatically and unconsciously in the mind, restructuring and reorganising items in information systems requires deliberate effort. This issue is discussed in detail in Section 3.4.

4. Categories in the mind do not have a simple one-to-one correspondence with the labels used to describe them. This means that the vocabulary we use to describe categories can be inconsistent and difficult to predict. Section 3.5 describes this issue and reviews work by information scientists identifying this problem.

Given these issues that emerge from the literature, what are the implications for the design of information systems? Here, I propose three implications:

1. When designing information systems the categorisation structures need to be carefully matched to the context in which the system is to be used. This goes beyond simply making sure the vocabulary of the information system matches that of its users, but extends to the information architecture as well.

2. All information systems require some mechanism for dealing with category shifts and changes. This is well recognised in the literature; however, most proposed solutions tend to assume the presence of a trained expert to manage the change. In many circumstances this is not always practical, so other measures must be in place to cope with category change.

3. Careful attention needs to be paid to the vocabulary of a categorisation system. Again, this is well recognised in the literature; however, the results of studies on user vocabularies suggest that many more aliases are required than most thesaurus systems provide.

These implications form part of the problem definition formalised in Chapter 5.

3.1 CONCEPTIONS OF CATEGORISATION

3.1.1 *Metaphorical Understanding and Categorisation*

Whenever we learn a new, abstract, concept we always conceptualise it in terms of things that we already understand. Because we experience everything through the sensory apparatus provided by our bodies, we tend to reason about abstract concepts in terms of concrete concepts we already understand from bodily experiences [72]. This projection from embodied experiences to abstract concepts is called *conceptual metaphor*, and forms the basis of much of our reasoning and thinking [71].

We also have the ability to combine conceptual metaphors to create more complex metaphors [72]. We may even use multiple, somewhat different, metaphors to understand a single concept. All of these complex metaphors, however, are built from very simple metaphors

based on bodily experiences. These building blocks are called *primary metaphors*.

Many of our primary metaphors come from the experience of being able to manipulate *objects*. Most of us have hands and arms that we can use to carry, push, pull, and otherwise exert force on objects. Our brains are much quicker at recognising objects at the level of things we can interact with [113, 121]. Our brains are hard-wired (so to speak) to work this way because the ability to recognise and manipulate objects (such as food, for example) is essential to our survival.

One common thing that we do with objects is to place them in containers.* In addition, we often group similar things together in containers. This forms the basis of the primary metaphor CATEGORIES ARE CONTAINERS [72]. For example, we say such things as 'I put it in the miscellaneous category' or 'I don't like to be put in a box'. Hence, when we think and reason about categorisation we often conceptualise it in terms of placing similar objects together in a bounded region of space.

Categorisation by placing objects in containers is something we do all the time in our daily lives. We place clothes in wardrobes, books on shelves, and food in cupboards. Kitchens are a prime example of this. Cutlery is customarily placed in a drawer, and the drawer is subdivided into containers for forks, knives and spoons; saucepans are placed with other saucepans and food is placed in cupboards with other food. We categorise dirty clothes by placing them in a basket and clean clothes by placing them on a shelf. It makes sense for us to conceptualise categorisation in this way because it is something we do naturally and unconsciously all the time.

3.1.2 *The Classical Theory of Categorisation*

The CATEGORIES ARE CONTAINERS metaphor is useful. It suits a great many situations and we use it without thinking. In fact, it is difficult to conceptualise categorisation in any other way. Because of this, people often think about categories not so much as mental constructs, but rather as things which exist in the external world. By this reasoning, when we learn categories we do not create them ourselves but rather learn to recognise categories that already exist [62].

Until research in cognitive science demonstrated graded category structures, this view of categories was taken to be the nature of re-

* Here, 'container' refers to any bounded region of space

ality, and 'the "right" way to think about categories, concepts, and classifications' [43, p. 340].

> 'The world of experience was assumed to consist of a set of predetermined categories, each defined by a set of essential features represented by a category label; and all members of a given category were assumed to share a set of essential features that was identified by the category label and could be apprehended by all members of the linguistic community' [62].

This assumption that categories can be defined by 'essential features' is the basis of the classical theory of categories. This understanding of categorisation rests on three propositions [quoted from 62, see also 122]:

I. The intension of a category is a summary representation of an entire category of entities.

II. The essential features that comprise the intension of a category are individually necessary and jointly sufficient to determine membership within the category.

III. If a category (A) is nested within the superordinate category (B), the features that define category (B) are contained within the set of features that define category (A)

According to proposition I, the *intension* of a category (its definition, or essence) is the set of essential features that a member must have to belong to the category. This set of features can then be used to represent the category as a whole. Because each member of the category must share these features (by definition) there cannot be any one member that is a better example of the category than any other. That is, the category structure is ungraded.

Proposition II means that any item that does not possess all the essential features that define the category cannot be a member of the category. Because of this, membership of a category is binary. That is, either an item belongs to the category or it does not; there are no degrees of membership. The boundaries of the category are fixed and well-defined. In terms of the container metaphor, either an object is inside the container or it is outside. It cannot be *somewhat* inside or *mostly* outside.

Proposition III defines an hierarchical structure amongst categories. This idea is an entailment of the container metaphor. If I have a small container, and place it inside a larger container, then all the objects in the smaller container are now also within the larger container. Projecting this to categorisation, by proposition I, items in a subordinate category must share all the essential features of the superordinate category.

The classical theory of categories is a powerful idea because it is (reasonably) simple and can be applied to a great many things. It reflects our reasoning based on the CATEGORIES ARE CONTAINERS metaphor, which in turn reflects much of our experience of the world. In spite of its usefulness however, we know intuitively that many things don't quite fit this view of the world. In addition, research into categorisation shows that the classical theory of categories does not reflect how we categorise in practice.

3.2 CATEGORY STRUCTURE IN THE MIND

Research into categorisation has shown that the classical theory of categorisation does not adequately account for the way we categorise. Researchers have demonstrated that not only do people seem to be unable to identify the essential features of categories [51, 112], but they have also demonstrated that categories appear to have both graded internal structure and fuzzy boundaries.

3.2.1 *Graded Structure*

Researchers have repeatedly demonstrated that some category members are generally considered more typical than others. For example, people will state that a *robin* is a better example of a *bird* than an *ostrich*, while a *pigeon* falls somewhere between the two. This kind of internal structure has also been demonstrated for a wide variety of common objects such as furniture, fruit, vehicles, weapons, sports, colours and shapes. [110, 111].

In addition to graded internal structure, some things are better non-members than others. A *chair* is less like a *bird* than a *butterfly*. Barsalou [10] conducted an experiment in which participants were given a set of six items and asked to group together those that belonged to a given category. Participants were then asked to rank all six items according to how *typical* they were of the given category. The results

showed that even when participants disagreed about whether particular items belonged to a category, there was a high degree of agreement on how to rank the items. This shows that the graded structure of categories extends beyond the category boundary.

3.2.2 *Fuzzy boundaries*

Continuing with the *bird* example, we note that some categories do have definite boundaries; butterflies and bats are not birds, while penguins and ostriches are. Many categories however, have fuzzy or ill-defined boundaries [88]. That is, members can *somewhat* belong to a category. In many cases the degree of membership can even be measured on a scale of zero to one. For example, a man of 215 cm (7′) would generally be considered a *tall man*; we could assign this man a degree of membership 1.0 in the *tall man* category. On the other hand, a man of 150 cm (4′ 11″) would generally not be considered a *tall man*; we could assign this man a degree of membership 0.0. A man of 180 cm (6′ 1″), however, would fit somewhere in between. He is a *tallish* man.

3.2.3 *Basic Level Categories*

There is also much evidence to suggest that not all categories are created equal. Some categories appear to be more *basic* and more easily recognised than others. Brown [21] observed that an object can belong to many different categories:

> The dime in my pocket is not only a *dime*. It is also *money*, a *metal object*, a *thing*, and, moving to subordinates, it is a *1952 dime*, in fact a *particular 1952 dime* with a unique pattern of scratches, discolorations, and smooth places. [...] The dog out on the lawn is not only a *dog* but is also a *boxer*, a *quadruped*, an *animate being*; it is the *landlord's dog*, named *Prince*.

In spite of the many categories to which an object belongs, we tend to use one or two names most often. So, the dime is referred to commonly as *dime* or *money* more commonly than *metal object* or *particular 1952 dime*. The dog on the lawn is commonly referred to as *dog* or *Prince*, but less commonly as a *boxer*, *quadruped*, or *animate being*.

These are not just linguistic conventions, but reflect a basic level of cognitive recognition [94]. These basic level categories 'optimally

fit our bodily experiences of entities and certain extremely important differences in the natural world' [72]. For instance, it is much more important for us to distinguish a cow from a tiger than it is to distinguish between two different species of tiger. We are much quicker at recognising basic-level categories, and develop the ability to recognise basic-level categories at a much earlier age [113]. Basic-level distinctions are generally the most useful distinctions to make since they are at the level that we interact with things in the world.

Categorisation is more complicated than the CATEGORIES ARE CONTAINERS metaphor would imply. Categories can have graded internal structures and some feature fuzzy boundaries. Furthermore, some categorisations are cognitively easier to make than others because they are basic to our cognitive functions. Hence, the way we categorise in practice is quite different from the classical theory of categorisation.

3.3 CLASSIFICATION VERSUS CATEGORISATION

When we reason about categorisation we still tend to use the CATEGORIES ARE CONTAINERS metaphor because it is so intuitive, and generally works quite well for basic-level categorisations. Because the metaphor works so well, then there is often little cause to question it, or even notice its existence. When it comes to categorising information, however, this way of thinking can cause problems.

One problem in particular is confusion over the difference between classification and categorisation. In modern usage, most people tend to use the terms interchangeably. Indeed, if the classical theory of categorisation were the only correct way to view the world, then there would be very little distinction between the two. There are important differences, however, which have implications for the way we organise information.

3.3.1 *Classification*

Jacob [62] defines *classification* as a process that involves 'the orderly and systematic assignment of each entity to one and only one class within a system of mutually exclusive and non-overlapping classes.' Similarly, a *classification scheme* is 'a set of mutually exclusive and non-overlapping classes arranged within a hierarchical structure and reflecting a predetermined ordering of reality.'

These definitions show a strong similarity to the classical theory of categories. An item is placed in a classification scheme according to a set of predetermined rules. If the set of rules is not met for a particular class then it cannot belong to that class. This has two implications:

1. CLASS BOUNDARIES ARE FIXED. An item either does, or does not fulfil the predetermined rules for membership in a class; so we can say that classes have binary membership, either in or out.

2. CLASSES DO NOT HAVE GRADED STRUCTURE. All members of the class must fulfil the requirements for membership, hence all members represent the class equally well.

In a classification scheme, classes are also mutually exclusive and non-overlapping. This means that any entity can only belong to one *correct* class.

The systematic properties of classification schemes can be extremely useful. The predetermined rules of a classification scheme encode knowledge about the class members. Jacob [62] illustrates this through the example of taxonomy schemes used to label plants and animals:

> Each class in the taxonomic scheme is given a unique name that is used to refer to all entities that display the complete set of features defining the class. And, because it is universally employed to identify all members of a given class, this label provides access to the accumulated knowledge about those entities, not as individuals but as members of a particular class. [...] Using the taxonomic name, a member of a biological class is recognizable wherever it occurs, regardless of natural language or the local name(s) by which it may be known.

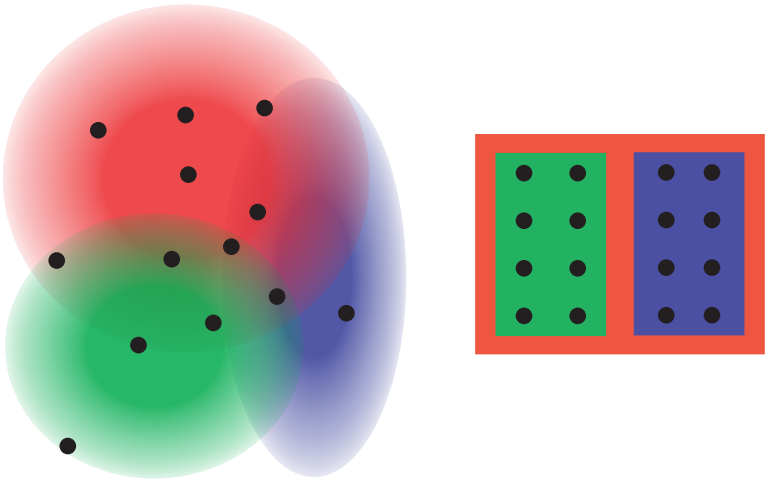
Knowing that Bob is a *goldfish* provides me with information about Bob. I know that goldfishes are members of the class *fish* and that fish have scales and live in water. Goldfish are also members of the class *freshwater fish*, which tells me even more about where they live. Each level of the taxonomy encodes different information that I can know about Bob.

Another good example of a classification scheme is the periodic table of elements. Knowing the atomic number of an element encodes certain information about its atomic structure. I can also infer information from the superordinate categories. For example, if I know that a particular substance is Helium (atomic number 2), then I also know

from the classification scheme that it is a noble gas which has certain chemical properties.

3.3.2 *Categorising and Classifying Information*

In contrast to classification, categorisation is simply grouping things together for a particular purpose. In terms of cognition, 'categories are groups of distinct abstract or concrete items that the cognitive system treats as equivalent for some purpose' [84]. As we have seen, categories may have graded structure, fuzzy boundaries, and need not be mutually exclusive. This is illustrated in Figure 6. The differences between categorisation and classification are summarised in Table 4.



(a) Categorisation (b) Classification

Figure 6: Categories have graded membership, fuzzy boundaries and are not mutually exclusive.

When it comes to organising information, historically, librarians and information scientists have developed classification schemes to make things easier to find later. The physical constraints of books, shelves and library buildings matched well with the requirements for a classification scheme. A book can only be on one shelf at one time. A classification scheme such as the Dewey Decimal System or the Library of Congress Subject Headings (LCSH), provides a set of rules for determining which single shelf the book belongs on. Thus, if I know the classification system, I should be able to find exactly where a particular book should be shelved.

As discussed in Chapters 1 and 2, organising information is problematic. A book or document is generally classified according to what

Table 4: Comparison of Categorisation and Classification [taken from 62].

CATEGORISATION	CLASSIFICATION
Process	
Creative synthesis of entities based on context or perceived similarity	Systematic arrangement of entities based on analysis of necessary and sufficient characteristics
Boundaries	
Because membership in any group is non-binding, boundaries are 'fuzzy'	Because classes are mutually-exclusive and non-overlapping, boundaries are fixed
Membership	
Flexible: category membership is based on generalised knowledge and/or immediate context	Rigourous: an entity either *is* or *is not* a member of a particular class based on the intension of a class
Criteria for Assignment	
Criteria both context-dependent and context-independent	Criteria are predetermined guidelines or principles
Typicality	
Individual members can be rank-ordered by typicality (graded structure)	All members are equally representative (ungraded structure)
Structure	
Clusters of entities; may form hierarchical structure	Hierarchical structure of fixed classes

it is *about*. This means that a cataloguer has to read and comprehend the item to understand what it is *about* so that it can be classified [37]. But it is entirely possible for a book or document to be *about* any number of things. For example, I could write a book critiquing the forensic methods employed by Sir Arthur Conan Doyle's *Sherlock Holmes*, supplemented with explanations of how modern forensic scientists would approach similar cases. Now, is my paper about *forensic science* or *literature*? One could make a case for either. Since my essay can only be placed on one shelf however, the cataloguer must make a decision as to what the book is *really* about. Information (whether it be text, graphics, audio, or video) does not always fit well into classification schemes.

This is not to say that classification schemes are of no use. In many circumstances classification schemes provide powerful and effective tools for organising information. It all depends on the nature of the information and the purposes for which it is being organised. The important thing to note is that classification schemes are by no means the only way to organise information, and are not the same thing as categorisation.

3.4 SITUATED LEARNING AND CATEGORY DYNAMICS

Another property of categorisation at the cognitive level is that categories are highly contextual. That is, we learn and create categories to suit specific purposes in specific situations.

Our brains do not learn concepts in abstract. We always learn things in particular circumstances at particular times. It is impossible to separate what we have learned from how and why we learned it [20]. In many cases, what we learn is a by-product of some task we undertake, rather than its goal. The same is true for categorisation. As stated by Markman and Ross [84]:

> [P]eople's natural categories are acquired in the course of interacting with the categories and with category members. People classify novel items, make predictions about unknown properties, solve problems with categories, make explanations based on them, communicate about them, and form preferences. Each of these tasks leaves its mark on the representation of these categories. Often, category information is learned as a by-product of interactions with the category and is not the central goal of the interaction.

Another point to note is that our category structures change as we interact with category members. Our brains continually update the categories we create so that they better suit our purposes and to accommodate newly acquired knowledge. For example, many people's concept of the category *camera* has changed in recent years with the introduction of *digital cameras*. The appearance of cameras and the ways we interact with them have been changed by the introduction of new kinds of cameras. Yet it has not caused most people too much trouble. We still recognise a digital camera as something that can be used to take photographs.

The categories we create vary in their degree of stability. At any point in time category structures will encode some context-specific and some non-context-specific information [62]. For example, if I am thinking about *things to take on a holiday*, then what I take will depend entirely on my holiday destination. A trip to the Pacific islands will entail very different category members than a tour of Antarctica. However, regardless of where I am going, the category of things I call *socks* (for example) remains largely the same.

The flexibility of our cognitive processes allows us to learn and respond appropriately as we encounter new situations and objects. Thus, we are able to create *ad hoc* categories, such as *things to take on a holiday* [10]. This can create problems when we attempt to organise information artifacts outside of our minds however. When retrieving information we have a need for stability so that we can find things again: when I place a book on a shelf, I normally expect it to stay there until I retrieve it again (unless somebody else moves it). It's location and my means of accessing it are stable—I know where to look for the book if I want it again. By contrast, many of the category structures in my mind are context-specific, goal oriented, and therefore unstable. Categories like *Papers relevant to the journal article I'm writing* are a good example.

This varying degree of stability manifests itself in the way people organise papers in an office. Studies have found that people 'like to group items spatially and often prefer to deal with information by creating physical piles of paper, rather than immediately categorizing it into specific folders' [82]. In addition to this, we also create more stable filing systems using shelves, filing cabinets and drawers to create more stable information stores. However, unless people make a deliberate effort to manage and maintain the temporary piles, offices quickly become messy. Retrieving relevant information then becomes much more difficult [80]. Categories in our brains are continually be-

ing created, refined and discarded; however, we rarely have similar mechanisms to reflect this in the physical organisation of our offices. Piles of paper reflect unstable, ad hoc categories, yet creating, updating and maintaining them requires deliberate effort.

The same can be said of hierarchical folder schemes used in desktop computer systems. Desktop computers allow us to store data using interfaces that are built around physical metaphors for storing information. The Windows and Macintosh operating systems, for example, base the organisation of data around a *desktop* metaphor. The system allows us to store our *files* in *folders* which can also contain *subfolders*. This creates a hierarchical organisation where each file has a single location. Unfortunately, user interfaces based on the *desktop* metaphor tend to reflect the same difficulties in keeping information organised. Folders are often created to support a particular task, but the contents and name of the folders do not change when the task changes or is completed. Keeping an electronic file system organised can require just as much effort as keeping a paper file system organised.

Any system for organising information must balance the conflicting requirements for *stability* and *flexibility* [125]. The degree of flexibility and stability required will vary depending on the nature of the information system. Large libraries, for instance, require a high degree of stability, as they are used by many different people for a myriad of purposes. A system for organising personal internet bookmarks (such as those described in Section 6.6, on the other hand, require a much greater degree of flexibility.

This is much easier said than done however. Consider the SIMPRESS system described in Chapter 2. Imagine that a change in production techniques lead to a new class of problems occurring which had no appropriate category in the *concern type* list. Either all problems of this kind will be classified as *Other*, or a new category needs to be added to the system. This is relatively simple to do using the administrative interfaces provided. The problem is, however, that the need for a new category would not have become apparent until a significant number of entries had already been entered into the system; presumably categorised as *Other*. To facilitate retrieval and re-use, these entries should be re-categorised, requiring somebody to examine all of the entries in the *Other* category looking for problems of this kind. In such a time pressured environment however, this kind of task is not likely to be given a high priority (see Section 2.4). As observed by Malone [80], keeping information organised well is a difficult task.

3.5 THE VOCABULARY PROBLEM

The problem of unstable and subjective categorisations is even further exacerbated by the issue of language and vocabulary. The relationship between categories and the labels we use to describe them is complex. Given a set of category members, it is very difficult to predict what labels people will choose to describe the category.

As mentioned earlier, Brown [21] observed that when naming an object we can choose from a number of different names. The *tree* outside my window is also a *eucalypt*, a *plant*, and a *gum tree*. While I may use some of these terms more frequently than others, they are all valid ways of naming the tree outside my window.

This suggests that naming a thing is not the same as recognising a thing. However, the two are not unrelated. In order to name something I must also recognise it. Malt et al. [81] suggest that the relationship between the two acts, recognition and naming, is not simple:

> Despite their close connection, [...] the act of naming differs critically from the act of recognition. Naming is part of a communication process, whereas recognition is not. The name selected for an object may reflect requirements for successful communication, whereas the representation of commonalities presumably is influenced primarily by constraints such as storage efficiency and the ability to support inference. Because of this fundamental difference in the nature of the two acts, the coupling between recognition and naming may be less than perfect. In particular, we posit that names used for objects reflect influences that are independent of the process of internal representation.

Thus, categorisation is not simply a matter of assigning an object a category label. The category labels we use depend on purpose and context just as much as the category itself. In the mind, this contextual information is closely associated with the category. When it comes to information systems however, categories, labels and indexes are often intended to be utilised by many people in a variety of different contexts. This makes the process of choosing a vocabulary to describe information items rather difficult.

This issue of vocabulary is central to the organisation of information. Not only do people's categories differ, but their vocabulary for describing them differs also. Buckland [22] notes that in most information systems, there are up to five different vocabularies in use:

1. The vocabulary (or vocabularies) of the authors(s) of the documents searched;
2. The vocabulary of the cataloguer, used in creating representations of the documents, modifies, replaces, and/or supplements the author's vocabulary;
3. "See," "See also," and other syndetic structures modify, replace or supplement the cataloguer's vocabulary;
4. The vocabulary of the searcher; and
5. The vocabulary of the searcher as formulated as a search query.

In transferring between any of these vocabularies, there is the potential for mismatch. The searcher may be looking for A, while the author called it B, and the cataloguer interpreted it as C. Each of the different vocabularies is affected by the different purposes and contexts of each actor.

The difficulty this causes has been observed by a number of authors. Furnas et al. [42] studied the problem of naming commands for text-based user interfaces. They asked participants to choose likely words that could be used as commands for various (hypothetical) computer programs. In every case they found that the likelihood of two people favouring the same term was less than 0.2.

Collantes [27] conducted a similar study, but instead asked participants to give names for a variety of stimuli including pictorial symbols and abstracts from books. They then compared the degree of agreement between different participants, and agreement between the participants and LCSH. In this study, participants were allowed to give more than one name to each stimulus. They found that 'there is little agreement across people in the names they use to describe texts or illustrations. Likewise, there is little agreement in the names people use and the names recommended for use by LCSH'. They then conclude:

> These findings about the nature of names lead us to conclude that naming objects and concepts is not at all simple or straightforward. How people perceive and interpret objects and concepts affect the way they create names or search statements. If naming is a highly individualized process, then prediction is difficult.

Interestingly, Collantes [27] also found that participants 'did not automatically choose or use any term from the texts', indicating that

even full-text search indexes do not provide a guarantee of relevant names.

3.6 IMPLICATIONS

What then are the implications for the design of information systems? Here, I propose three:

1. Category structure needs to be carefully matched to the context in which the system is to be used.
2. All information systems require some mechanism for dealing with category shifts and changes.
3. Careful attention needs to be paid to the vocabulary of a categorisation scheme.

3.6.1 *Category Architecture*

As discussed in Sections 3.2 and 3.4, not all categories are the same. Categories in the mind vary in level of complexity and stability. Likewise, information systems deal with differing levels of complexity and stability. A hierarchical classification scheme will not be well suited to situations where categories change rapidly or cannot be prescribed by a predetermined set of features. Conversely, where categories are relatively stable, a classification scheme can be a powerful tool for organising information and encoding knowledge.

The idea that category architecture must be carefully matched with context is not a new one. However, the focus has often been on vocabulary rather than on structure. In recent times, however, the success of internet search engines has lead to the reexamination of categorisation structures. The field of Information Architecture has emerged in response to this shift [99]. Until recently, however, most information architecture seemed to be variations on either full-text search, or hierarchical classification.

3.6.2 *Category Dynamics*

Section 3.4 explored the issues of situated learning and category dynamics. Category structures in the mind are dynamic; continually updated and re-organised to accommodate new knowledge and understanding. While this happens somewhat automatically and uncon-

sciously in the mind, restructuring and reorganising items in information systems requires deliberate effort. Hence, information systems require mechanisms for dealing with these category shifts and changes. This is well recognised in the literature [125, for example]; however, most proposed solutions tend to assume the presence of a trained librarian. Even the literature on automatic classification assumes the presence of an expert able to train the auto-classifier and correct misclassifications.

Of course, if a highly-trained individual (or a team) is dedicated to the task of maintaining, reviewing and updating an information system, then categorisation will generally be performed well. Hence, libraries with dedicated librarians tend to be well organised. In many situations, however, a dedicated librarian is not available to maintain the categorisation system. This thesis explores methods of categorisation that do not rely on a single individual dedicated to the task of maintaining good organisation in the information system.

3.6.3 *Vocabulary*

Categories in the mind do not have a simple one-to-one correspondence with the labels used to describe them. This means that the vocabulary we use to describe categories can be inconsistent and difficult to predict. Thus, careful attention needs to be paid to the vocabulary of a categorisation system. Again, this is well recognised in the literature. Ontologies and enterprise thesauri are examples of attempts to address this problem. However, the findings of Furnas et al. [42] and Collantes [27] suggest that people use a vast array of terms to describe categories and that the small number of aliases usually provided by such systems are generally inadequate.

3.7 SUMMARY

We have (very quickly) reviewed the work of cognitive scientists in understanding the structure of categories in the mind, and investigated how this can cause problems when it comes to categorising information. Understanding these problems allows us to better evaluate actions which attempt to address them. Continuing on from this work, Chapter 4 examines the impact of a user interface that allows graded category membership on the degree of agreement between people categorising.

4

GRADED CATEGORISATION AND USER INTERFACES

In the previous chapter we explored categorisation in the mind, and described the graded structure of categories. In this chapter, we extend these concepts to user interfaces for categorisation. In particular, we investigate the question: *To what extent does an interface that allows graded membership improve accuracy of categorisation?* To answer this question, I conducted an experiment asking participants to categorise ten items containing complex information. Half the participants used a slider bar interface allowing graded categorisation, while the other half used a radio button interface allowing only boolean category membership. The results showed that using the graded interface increased accuracy by around 3%; however, people using the graded interface took, on average, 8.8 seconds longer to categorise. The results also showed that individual people were not particularly good at predicting the categorisations of their peers. However, there is potential to use contributions by several people as a prediction of group consensus.

4.1 INTRODUCTION

In Chapter 3 we saw that category structures in the mind are dynamic; continually being updated as new knowledge about category members and purposes becomes available. Thus, categorisation mechanisms for information systems must be flexible. At the same time, information systems must balance flexibility with stability so that information can be retrieved effectively.

The balance between these two requirements is largely determined by purpose and context. Ad hoc categories, such as *photographs from business trips* or *files for tomorrow's meeting* will change much more rapidly than, say, the LCSH. The number of people needing to use the categorisation system, the purpose of the categories, and the nature of category members, all influence category dynamics.

In many cases, the best solution for coping with category shift is to employ someone (such as a librarian or metadata expert) to manage the categorisation system—both to categorise new entries and to add, remove, or modify categories. In many cases, however, employing somebody to perform this task, even on a part-time basis, is not

practical.

Where no dedicated librarian or other expert is available, three alternatives are available for categorisation:

A. Attempt to classify the items using an automatic text classification algorithm.

B. Allow content authors to categorise their own entries.

C. Allow system users to categorise entries as they use them.

Each of these options has a number of disadvantages. Option (A) usually requires that someone is available to train the system and correct misclassified entries. Hence it does not really solve the problem of no expert being available.

Option (B) has the advantage that authors are likely to know the most about the content and thus have a good idea of how to categorise it. On the other hand, there is potential for authors to misuse categorisation to suit their own ends, creating the perception that author-contributed metadata lacks credibility. Thomas and Griffin [135] argue that there is little incentive for authors to put time and effort into providing quality metadata unless it is for the purposes of self-promotion. Thus, author generated metadata is perceived as inaccurate at best, and deliberately misleading at worst [33, 41].

Option (C) has the disadvantage that the users of a system are likely to have the least amount of expert knowledge either of how to categorise (as with the metadata expert), or what they are categorising (as with the content author). Hence, user-contributed metadata can be messy and inaccurate.

In spite of this, Surowiecki [132] suggests that the collective contributions of users can be aggregated to counteract individual error. Hence, in systems with multiple users, collaborative categorisation may be a viable alternative. In this chapter we explore the use of a slider-bar interface as a potential tool for improving the accuracy of categorisations provided by users.

4.1.1 *Graded Categorisation*

As discussed in Chapter 3, psychologists have demonstrated that categories in the mind have a graded structure, with some members being more typical than others [10, 111, 112]. By contrast, categorisation systems often require users to *classify* items—assigning them to non-graded, mutually-exclusive categories. It seems reasonable to

suppose then, that providing a graded categorisation interface would make categorisation more accurate and intuitive. Thus, I conducted an experiment to test two hypotheses:

Hypothesis 4.1 *Use of a graded categorisation interface will increase accuracy of categorisation*

Hypothesis 4.2 *Use of a graded categorisation interface will make categorisation more intuitive.*

To test these hypotheses, I conducted an experiment asking participants to categorise information artifacts using two different categorisation interfaces: one graded, the other ungraded.

4.2 METHOD

To conduct the experiment, I asked participants to play an internet-based game called 'PopCat' (short for Popular Categorisation). In groups of around 4–8 players, participants were shown a number of information artifacts and asked to categorise each using one of the two categorisation interfaces. After categorising the item, participants' answers were compared to others in the same group. For each artefact categorised, participants were given a score based on how close their categorisation matched the group mean.

The format of a game was chosen to encourage active participation. The competitive nature of a game gave participants an incentive to think carefully about their categorisation choices—to think 'How would others categorise this?' In doing so, participants were not simply stating a personal opinion, but making a prediction about how others would categorise the same item.

I recruited participants from students enrolled in a first-year engineering class offered by the Department of Engineering, ANU. Before the experiment sessions, participants were given an information sheet (see Appendix B) explaining the purpose of the study and giving instructions on how to play. I also explained to students that participation in the experiment was voluntary and would not affect their university course marks in any way.

Conducting the experiment sessions in computer labs at the university allowed each participant to have their own workstation and web-browser. I instructed participants to organise themselves into groups of four to six players, and to seat themselves close to their groupmates. Having the groups sit together made giving instructions easier for the instructor; however, physical proximity of group members

was not necessary to play the game. Early trials of the game were successfully conducted with participants communicating via instant messaging.

4.2.1 *Game Procedure*

The procedure for the game was as follows:

1. Participants were asked to 'log on' to the game system through the browser and to create a pseudonym to represent themselves in the game.
2. The PopCat system asked a few short survey questions to gather demographic information.
3. Once participants had completed the survey, the system presented a screen giving the option of creating a new game or joining an already 'open' game.
4. One participant from each group 'created' a new game and invited other participants from the group to join. When the game was created, the PopCat system randomly determined whether the group would use the Graded Interface (GI) or Non-Graded Interface (NGI) for categorisation. The system then showed a 'waiting' screen while participants waited for others to join the game.
5. Once all participants in a group had joined, the game initiator clicked a 'begin game' button, and all participants were then shown the categorisation interface (Figure 7).
6. Participants categorised the information artefact using one of the two interfaces (Figure 8), then clicked the 'submit' button.
7. The system waited until all participants had submitted their categorisation choices before calculating running scores and presenting the next artefact.
8. This process continued for ten rounds, after which a final score sheet was presented announcing the winner of the game.

For each artefact, participants submitted nine *votes* on the category degree of membership (DoM)—one vote for each category. For the NGI these variables could only take on the values *zero* or *one*, corresponding to 'Definitely No' or 'Definitely Yes' as shown in Figure 8. For the

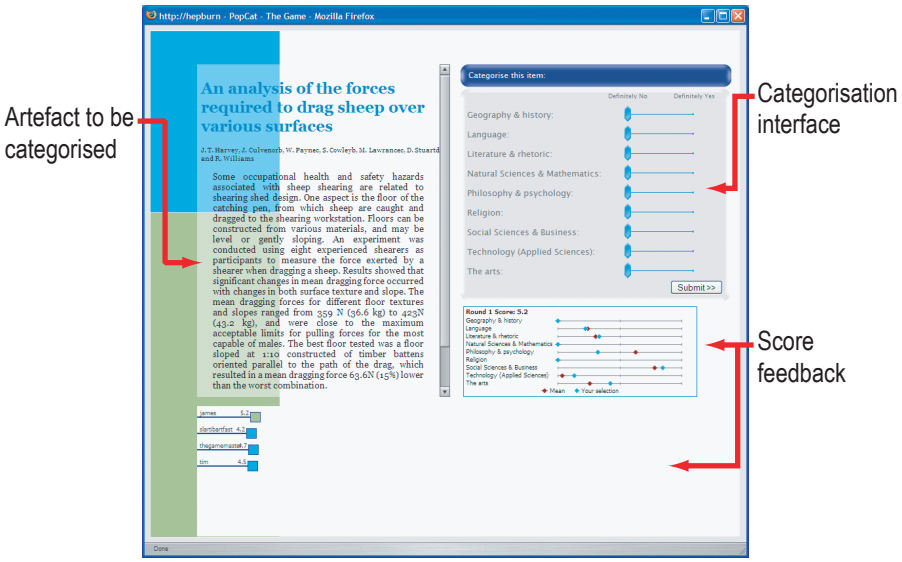
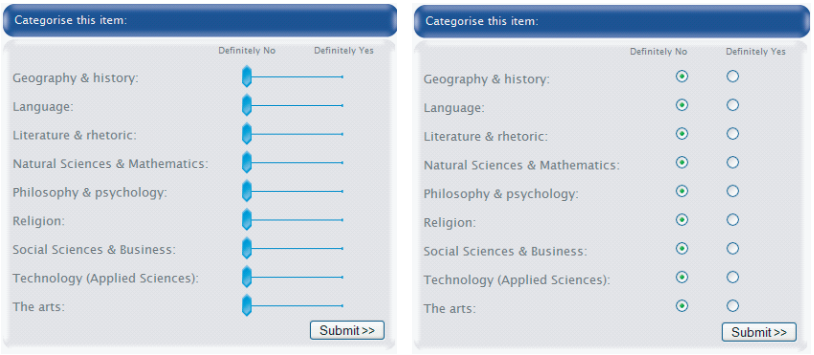


Figure 7: The user interface for the categorisation game.



(a) (b)

Figure 8: The two categorisation interfaces: (a) The graded interface (GI); (b) the non-graded interface (NGI).

GI these variables could take any value between zero and one, with a resolution of 0.01. These DoM variables were used to measure the in-game scores, and form the basis of the analysis in Section 4.3.

The system calculated in-game scores by determining how much each participant's categorisation choices differed from the mean of all players in their group. That is, the system calculated a score, $S_{i,j}$, for each person using the following equation:

$$S_{i,j} = \frac{10\sqrt{K}}{\sqrt{\sum_k (d_{i,j,k} - \overline{d}_{j,k})}} \tag{4.1}$$

where $S_{i,j}$ is the score given to player i for round j, K is the number of categories, $d_{i,j,k}$ is the DoM assigned to artefact j for category k by player i, and $\overline{d}_{j,k}$ is the mean DoM assigned to artefact j for category k by all members in the group. The cumulative score was displayed at the bottom of the screen as shown in Figure 7.

4.2.2 *Category Selection*

The categories used in the game were nine of the ten top-level categories specified by the Dewey Decimal System as shown in Figure 8. The first category of the Dewey Decimal System, Generalities, was excluded on the basis that it may serve as a kind of 'Miscellaneous' category into which any of the artifacts might reasonably be placed. The participants were not told that the categories were taken from the Dewey Decimal System, and it is unlikely that many of the participants would be intimately familiar with the subcategories excluded by removing the Generalities category.

4.2.3 *Selection of information artifacts*

Unlike other studies [eg. 27, 81, 111, 113], this experiment asked participants to categorise information artifacts, rather than simple objects or pictures. In addition, artifacts were selected to cover a broad range of subject areas and different media types (summarised in Table 5). They included academic abstracts, comic strips, television advertisements, and a poem.

Academic abstracts formed half of the artifacts to be categorised. An abstract is intended to summarise the entire paper and give the reader enough information to know what it is about, making it a good

Table 5: Artefact Media Types

MEDIA TYPE	NO. ARTIFACTS
Academic abstract	5
Comic strip	2
Television advertisement	2
Poem	1
Total	10

example of information categorisation.

As an information artefact, comic strips are of particular interest because of the way they combine text and graphics:

- They are normally stored in a raster image format (eg. bitmap, JPEG, or GIF) which means that although they may contain textual data, it is not usually in a form easily readable by a machine.
- Although a comic strip may contain textual information, other graphical elements are usually essential to understanding the comic.

Television advertisements were included as another form of media that may be found in a knowledge management system. They were also selected to be interesting in order to help keep participants active and interested in the game. The poem artefact was a sonnet by Shakespeare, and was included as an example of an item that may be outside of an Engineering student's domain of expertise.

4.2.4 *Participants*

As mentioned previously, participants were recruited from students enrolled in a first-year engineering class offered by the Department of Engineering at the ANU. Of the 126 participants who took part in the experiment, 61 used the NGI, while 65 used the GI. Table 6 summarises other characteristics of the participants.

Table 6: Participants

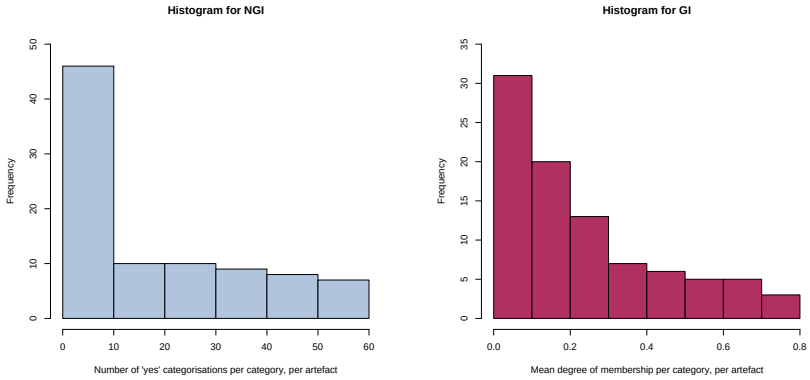
	FEMALE	MALE	TOTAL
English as a Second Language	3	30	33
English First Language	14	79	93
Total	17	109	126

4.3 RESULTS

The game system recorded two main variables for each question: 1) the nine DoM votes for each category; and 2) time taken to read (or view) and categorise the artefact. We examine each of these two variables in turn.

4.3.1 *Categorisation Results*

Tables 7–16 show the categorisation results for each artefact. On average, participants using the NGI voted 'yes' for 2–3 categories per artefact. Since this leaves 6–7 other categories as 'no', the majority of votes recorded were 'no' votes. The GI showed similar low DoM values. These trends are illustrated in Figure 9.



(a) NGI Histogram (b) GI Histogram

Figure 9: Histograms for each interface

Interestingly, no artefact received a unanimous positive vote for any category. The highest number of 'yes' votes for the NGI was 60 (98.4%), while the highest mean DoM using the GI was 0.784. This suggests that participants were more confident at stating an artifact *definitely did not* belong to a category, than they were at stating an arte-

fact definitely did belong to a category. This may be a result of the very broad categories selected for the game. The Dewey Decimal categories are intended to cover the whole gamut of human knowledge, and thus the top level categories necessarily combine wide-ranging fields of expertise. Thus, it would be difficult to find a prototypical example for any of the categories listed.

Table 7: Categorisation Results: Artefact 1

	NON-GRADED			GRADED	
Category	# Yes	# No	Ratio	Mean DoM	Propn.
Geography & History	0	60	0.000	0.028	0.006
Language	28	32	0.467	0.259	0.077
Literature & Rhetoric	30	30	0.500	0.316	0.112
Natural Sciences & Mathematics	2	58	0.033	0.081	0.021
Philosophy & Psychology	34	26	0.567	0.537	0.111
Religion	2	58	0.033	0.047	0.018
Social Sciences & Business	41	19	0.683	0.712	0.091
Technology (Applied Sciences)	17	43	0.283	0.225	0.096
The Arts	17	43	0.283	0.149	0.061

Table 8: Categorisation Results: Artefact 2

	NON-GRADED			GRADED	
Category	# Yes	# No	Ratio	Mean DoM	Propn.
Geography & History	9	52	0.148	0.071	0.013
Language	5	56	0.082	0.081	0.034
Literature & Rhetoric	1	60	0.016	0.083	0.039
Natural Sciences & Mathematics	52	9	0.852	0.646	0.062
Philosophy & Psychology	3	58	0.049	0.104	0.035
Religion	0	61	0.000	0.019	0.006
Social Sciences & Business	22	39	0.361	0.291	0.085
Technology (Applied Sciences)	53	8	0.869	0.645	0.092
The Arts	0	61	0.000	0.027	0.006

Table 9: Categorisation Results: Artefact 3

	NON-GRADED			GRADED	
Category	# Yes	# No	Ratio	Mean DoM	Propn.
Geography & History	0	61	0.000	0.024	0.003
Language	4	57	0.066	0.087	0.027
Literature & Rhetoric	1	60	0.016	0.076	0.018
Natural Sciences & Mathematics	42	19	0.689	0.442	0.105
Philosophy & Psychology	54	7	0.885	0.606	0.096
Religion	2	59	0.033	0.059	0.025
Social Sciences & Business	21	40	0.344	0.375	0.097
Technology (Applied Sciences)	21	40	0.344	0.280	0.072
The Arts	3	58	0.049	0.082	0.022

Table 10: Categorisation Results: Artefact 4

	NON-GRADED			GRADED	
Category	# Yes	# No	Ratio	Mean DoM	Propn.
Geography & History	1	60	0.016	0.024	0.003
Language	35	26	0.574	0.381	0.075
Literature & Rhetoric	42	19	0.689	0.480	0.102
Natural Sciences & Mathematics	1	60	0.016	0.032	0.009
Philosophy & Psychology	26	35	0.426	0.302	0.075
Religion	0	61	0.000	0.054	0.022
Social Sciences & Business	15	46	0.246	0.197	0.049
Technology (Applied Sciences)	10	51	0.164	0.122	0.034
The Arts	31	30	0.508	0.337	0.088

Table 11: Categorisation Results: Artefact 5

	NON-GRADED			GRADED	
Category	# Yes	# No	Ratio	Mean DoM	Propn.
Geography & History	7	54	0.115	0.135	0.046
Language	51	10	0.836	0.687	0.089
Literature & Rhetoric	55	6	0.902	0.722	0.094
Natural Sciences & Mathematics	0	61	0.000	0.024	0.004
Philosophy & Psychology	14	47	0.230	0.249	0.051
Religion	10	51	0.164	0.115	0.033
Social Sciences & Business	2	59	0.033	0.075	0.016
Technology (Applied Sciences)	0	61	0.000	0.016	0.004
The Arts	48	13	0.787	0.577	0.117

Table 12: Categorisation Results: Artefact 6

	NON-GRADED			GRADED	
Category	# Yes	# No	Ratio	Mean DoM	Propn.
Geography & History	3	58	0.049	0.062	0.013
Language	3	58	0.049	0.133	0.041
Literature & Rhetoric	9	52	0.148	0.189	0.073
Natural Sciences & Mathematics	50	11	0.820	0.520	0.100
Philosophy & Psychology	36	25	0.590	0.389	0.096
Religion	0	61	0.000	0.050	0.013
Social Sciences & Business	5	56	0.082	0.146	0.045
Technology (Applied Sciences)	34	27	0.557	0.372	0.097
The Arts	2	59	0.033	0.111	0.036

Table 13: Categorisation Results: Artefact 7

	NON-GRADED			GRADED	
Category	# Yes	# No	Ratio	Mean DoM	Propn.
Geography & History	1	60	0.016	0.080	0.039
Language	5	56	0.082	0.100	0.036
Literature & Rhetoric	3	58	0.049	0.097	0.038
Natural Sciences & Mathematics	11	50	0.180	0.273	0.073
Philosophy & Psychology	14	47	0.230	0.173	0.056
Religion	1	60	0.016	0.040	0.024
Social Sciences & Business	35	26	0.574	0.528	0.099
Technology (Applied Sciences)	60	1	0.984	0.784	0.048
The Arts	39	22	0.639	0.295	0.079

Table 14: Categorisation Results: Artefact 8

	NON-GRADED			GRADED	
Category	# Yes	# No	Ratio	Mean DoM	Propn.
Geography & History	2	59	0.033	0.064	0.012
Language	18	43	0.295	0.281	0.068
Literature & Rhetoric	18	43	0.295	0.168	0.034
Natural Sciences & Mathematics	20	41	0.328	0.226	0.050
Philosophy & Psychology	57	4	0.934	0.676	0.074
Religion	0	61	0.000	0.035	0.008
Social Sciences & Business	39	22	0.639	0.484	0.087
Technology (Applied Sciences)	7	54	0.115	0.157	0.044
The Arts	2	59	0.033	0.131	0.033

Table 15: Categorisation Results: Artefact 9

	NON-GRADED			GRADED	
Category	# Yes	# No	Ratio	Mean DoM	Propn.
Geography & History	10	51	0.164	0.164	0.057
Language	9	52	0.148	0.160	0.023
Literature & Rhetoric	13	48	0.213	0.164	0.044
Natural Sciences & Mathematics	2	59	0.033	0.059	0.011
Philosophy & Psychology	36	25	0.590	0.419	0.073
Religion	2	59	0.033	0.089	0.035
Social Sciences & Business	43	18	0.705	0.492	0.093
Technology (Applied Sciences)	23	38	0.377	0.225	0.050
The Arts	23	38	0.377	0.234	0.089

Table 16: Categorisation Results: Artefact 10

	NON-GRADED			GRADED	
Category	# Yes	# No	Ratio	Mean DoM	Propn.
Geography & History	50	11	0.820	0.587	0.108
Language	6	55	0.098	0.141	0.032
Literature & Rhetoric	4	57	0.066	0.107	0.029
Natural Sciences & Mathematics	24	37	0.393	0.242	0.049
Philosophy & Psychology	7	54	0.115	0.181	0.044
Religion	2	59	0.033	0.075	0.022
Social Sciences & Business	47	14	0.770	0.442	0.073
Technology (Applied Sciences)	22	39	0.361	0.228	0.041
The Arts	2	59	0.033	0.092	0.032

As one would expect, there was a strong correlation between the proportion of 'yes' votes for the NGI and mean DoM for the GI (R^2 = 0.929). Figure 10 shows a scatter plot of proportion of 'yes' votes versus mean DoM. Interestingly however, the slope of the regression line is not 1.0 but 0.66; suggesting that participants using the GI tended to understate category membership.

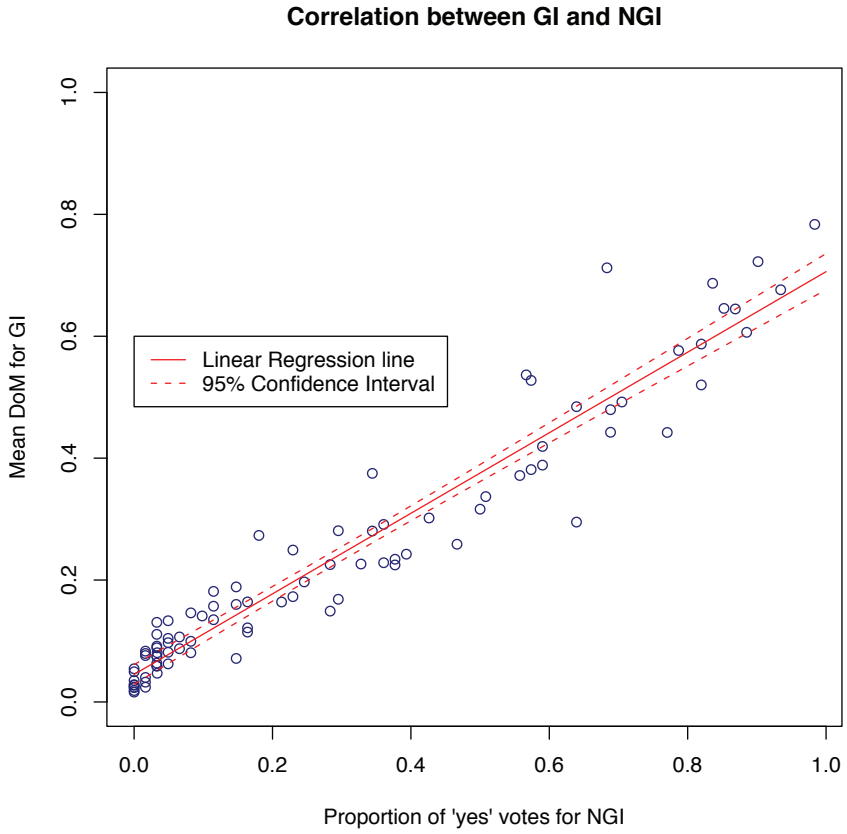


Figure 10: Correlation between GI and NGI

Looking just at the GI, we can examine how much participants agreed with one another by examining the variance in DoM assignments. By plotting variance against mean DoM (Figure 11) we can see that participants tended to disagree more over mid-range categorisations than they did for more definite *in* or *out* examples.

Taking these two observations together highlights a problem when analysing DoM values. Since there is a strong correlation between the proportion of 'yes' votes and mean DoM, this indicates that agreement amongst participants and mean DoM are dependent variables. That is, when people disagree on membership (yes/no), they also disagree on DoM. Hence, when DoM values are in the mid ranges (around 0.3–0.6),

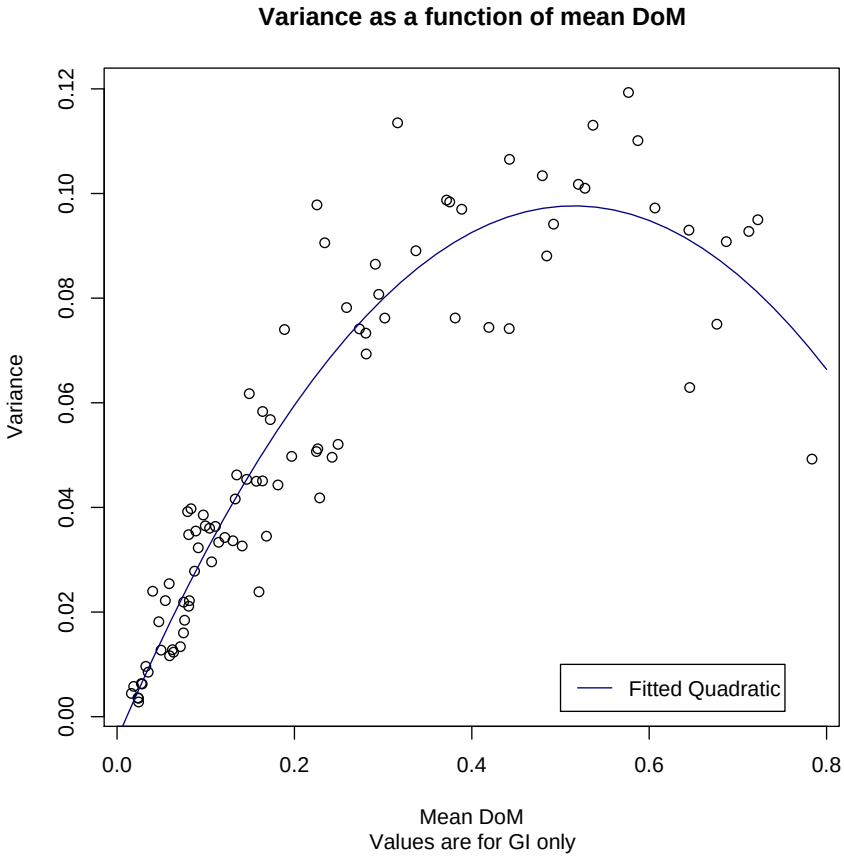


Figure 11: Variance for GI versus mean DoM

we cannot say whether the mean DoM is a general consensus or just a result of general disagreement.

4.3.2 *Categorisation Accuracy*

To investigate categorisation accuracy, we ask the question *how good were participants at predicting the judgements of their peers?* We measure this by taking the correlation between all pairs of DoM values for the same question, category and interface. That is, we take one person's DoM vote, and compare it with everyone else's vote for that question and category. This gives an indication of how good individual participants were at predicting what their peers would do.

For the GI, the overall correlation value was 0.43. For the NGI, the value was slightly lower, 0.41. While there is a definite correlation, these values are much less than 1.0, suggesting that individual participants were not particularly good at predicting their peers' categorisations.

These values are consistent with the findings of Barsalou [11], who measured similar correlation values for simple category structures among American university students. Surveying a range of experiments which asked participants to rank category members in order of typicality, Barsalou found average correlation values ranging from 0.30 to 0.60. Our values of 0.43 and 0.41 are well within this range. It is interesting that the NGI falls in this range however, since it does not measure graded structure, but only membership.

Also consistent with the findings of Barsalou [11], the results here show that homogenous groups did not significantly raise correlation values above the overall levels of 0.43 and 0.41 (Table 17).

Table 17: Pair-wise correlation values within homogenous groups

	GI	NGI
English First Language	0.46	0.43
English as a Second Language	0.38	0.35
Males	0.43	0.42
Females	0.47	0.36

For the purposes of organisation, we are not so much interested in an individual's ability to predict another individual's categorisation, but rather an individual's ability to predict the majority categorisation. To examine this, we measure the correlation between $d_{i,j,k}$ and

the mean DoM value assigned by all players *excluding* player i. To simplify our notation, we use $\hat{d}_{i,j,k}$ to denote this adjusted mean. Mathematically, we can describe $\hat{d}_{i,j,k}$ in terms of $\overline{d}_{j,k}$ (the mean for all players), as follows:

$$\hat{d}_{i,j,k} = \frac{P\overline{d}_{j,k} - d_{i,j,k}}{P - 1} \tag{4.2}$$

where P represents the total number of players.

Measuring the correlation between d and $\hat{d}$, we find much higher values: 0.65 for the GI and 0.64 for the NGI. Here, the difference in correlations is not statistically significant.

As with the pair-wise comparisons, factors such as language and sex did not significantly affect correlation values (Table 18). Note that in Table 18, $\hat{d}_{i,j,k}$ values are calculated for all players. For example, the value for females indicates the ability of an individual female to predict the overall mean, rather than the mean of all females.

Table 18: Correlation between $d_{i,j,k}$ and $\hat{d}_{i,j,k}$

	GI	NGI
English First Language	0.67	0.65
English as a Second Language	0.61	0.59
Males	0.65	0.64
Females	0.69	0.59

Correlation is one measure. Another question of interest is *How often did participants agree on a categorisation?* If we have an answer to this question, then we can estimate the probability that two participants would place an artefact in the same category.

To decide if two participants agreed for the NGI is relatively easy—either they both answered the same way, or they didn't. With the GI, however, the situation is slightly more complicated. To measure agreement for the GI we assume that the boundary for category membership is 0.5. We can then define an agreement function, δ, for the agreement between players a and b:

$$\delta(d_{a,j,k}, d_{b,j,k}) = \begin{cases} 1.0, & \text{if } d_{a,j,k} \geqslant 0.5 \text{ and } d_{b,j,k} \geqslant 0.5 \\ 1.0, & \text{if } d_{a,j,k} < 0.5 \text{ and } d_{b,j,k} < 0.5 \\ 0.0, & \text{otherwise.} \end{cases} \tag{4.3}$$

This equation works equally well for both interfaces.

A pair-wise comparison shows that participants agreed with each other 78.6% of the time using the GI and 75.7% of the time using the NGI. That is, agreement for the GI was 3% higher than participants using the NGI. While not enormous, the difference is statistically significant ($p < 2 \times 10^{-16}$). This is interesting because even when we remove the extra information collected by the GI by applying our δ function, the accuracy of the GI is still higher.

As with the correlation values, agreement was not drastically improved when measured within homogenous groups (Table 19).

Table 19: Agreement percentage within homogenous groups

	GI	NGI
English First Language	78.6	76.4
English as a Second Language	79.7	73.6
Males	78.2	75.6
Females	80.1	76.9

In both the pair-wise correlation measurement and the agreement measure, the GI shows a small, but statistically significant, improvement over the NGI. Hence we confirm our first hypothesis: the use of the GI does improve categorisation accuracy.

4.3.3 *Time To Categorise*

Our second hypothesis was that the GI is more intuitive than the NGI. To examine this, we look at the time taken to categorise each artefact. The game system recorded time taken from when the artefact was first displayed until the user clicked the "Submit" button (to an accuracy of 1/10 second). Thus, the recorded time was a measure of the time taken to read or view the artefact as well as the time taken to categorise the item. This does not allow a true measurement of how long the user took to categorise an artefact; however, it does allow comparison between the two interfaces since the artifacts were identical regardless of which interface was used.

Table 20 shows the mean time to categorise for each artefact. Due to the range of media types, the time taken to view/read each artefact varied a great deal. Artifacts 7 and 9, for instance, were videos which took some time to download and then display. In addition to this, a number of participants found the videos so amusing that they

Table 20: Mean time taken to categorise each artefact. Values are in seconds.

ARTEFACT	NGI	GI	DIFF.
1	89.4	121.4	31.9
2	108.8	123.7	14.8
3	81.4	97.1	15.7
4	62.2	75.2	13.0
5	66.9	71.9	5.0
6	77.9	93.1	15.1
7	205.5	180.8	-24.7
8	90.3	98.1	7.7
9	168.2	171.8	3.6
10	78.6	84.5	5.9
Mean	103.0	111.8	8.8

watched them multiple times. Hence, these artifacts are skewed towards longer categorisation times.

In spite of this variance between artifacts, however, there is still a consistent trend towards shorter categorisation times for the NGI. Only artefact 7 was inconsistent with this trend. Performing a t-test, the mean times to categorise for each interface are different with $p = 0.00913$. Thus, we conclude that the GI was slower to use than the NGI.

The difference in times does not mean however, that we can immediately conclude that the GI is less intuitive. First we must take into account the time required to physically interact with each interface. Assuming the participant moves each slider or radio-button, a GOMS analysis [109] gives a conservative estimate of 26.9 seconds for the NGI and 38.6 seconds for the GI. Assuming that participants took roughly the same amount of time to read/view each artefact, this more than accounts for the 8.8 second difference in mean time to categorise.

This reveals a draw-back to using slider-bars as the GI for this experiment. The slider bars result in more physical interaction with the interface. A better approach would have been to use a row of 5–8 radio buttons as shown in Figure 12. In this situation the physical interaction time for the interfaces would be much the same, allowing a more meaningful comparison of times between interfaces. Hence, we conclude that we can neither confirm nor reject Hypothesis 4.2.

Exploring the time to categorise further shows that first-language

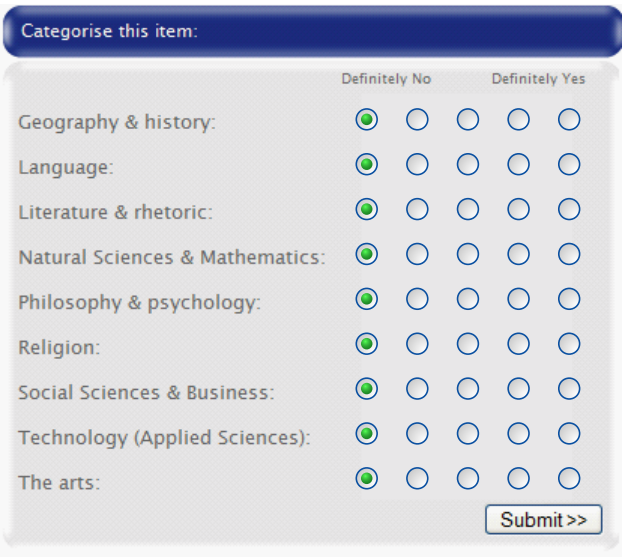


Figure 12: Alternative GI

also impacted time to categorise ($p = 7.8 \times 10^{-7}$) Participants with English as their first language were on average 32.1 seconds faster when using the NGI and 8.8 seconds faster when using the NGI, than were people with English as a second language. Surprisingly, people with English as a second language were quicker using the GI than the NGI (Table 21). This may be due to difficulties in comprehending the artifacts for participants with English as a second language. Thus, while there is a clear difference in categorisation time for people with English as a first language across the interfaces ($p = 0.0001427$), the difference for people with a English as a second language was not statistically significant ($p = 0.2104$). This would suggest that as the time taken to comprehend the artefact increases with language difficulties, the difference in time due to the interface becomes insignificant.

Table 21: Mean time to categorise by first language (seconds)

	NGI	GI
Eng. First Lang.	95.1	109.3
Non-Eng. First Lang.	127.2	118.1

Interestingly, there was no correlation between time taken to categorise and categorisation accuracy for the GI. Neither was there any correlation between time taken and agreement for either interface. It seems that the time taken was completely independent of participants' ability to predict their peers' categorisations.

4.4 DISCUSSION

The measurements in Section 4.3.2 showed that participants using the GI were better at predicting the categorisations of their peers. Hence, it seems reasonable to suggest that the GI captures more information than the NGI. In one sense, this is obvious since the NGI allows only yes/no values, while the GI recorded DoM to a resolution of 0.01. However, even when the DoM values produced by the GI were rounded to the nearest one or zero, it still gave a higher degree of agreement. So, I suspect that that the GI forced participants to think more carefully about their categorisations than the NGI did.

When it comes to information system design then, the choice of categorisation interface presents us with a familiar trade-off. The GI can give an increase in accuracy, but at the cost of interaction time. As with most things, the choice depends on requirements for the context. If the end result desired is simply a judgement of yes/no category membership then the relatively small increase in accuracy does not seem worth the added time. This is especially true where users are likely to be under pressure to categorise quickly (as in Chapter 2).

On the other hand, the GI captures more information than just yes/no categorisation judgements. A graded DoM value can be used to implement a user interface where category members are displayed and ranked according to their typicality, similar to the way search engines order results by relevance. This, in effect, creates a categorisation system based on fuzzy sets. The benefits of facilitating this kind of retrieval may outweigh the cost of extra time.

4.4.1 *Categorisation Accuracy*

In Section 4.3.2 we saw that participants were not particularly good at predicting the categorisation choices of other players. The pair-wise correlation values were only around 0.43, and correlation with $\hat{d}$ was only around 0.65. On average, an individual's categorisations agreed with roughly 75% of other people's categorisations. Even using the GI improved categorisation agreement only slightly. The categorisations of others are difficult for a single individual to predict.

The results from Section 4.2 however, showed that a group consensus does appear to exist when the mean all DoM values are taken. That is, participants did not disagree to the extent that the mean DoM for most categories was 0.5. Rather, there was a definite consensus that some artifacts belonged to some categories more than others.

Taking these two observations together, we can say that although there is a definite group consensus on category DoM, individuals on their own are not particularly good at predicting it. To put this another way, if we wish to gain a good estimate of mean DoM, we need to rely on more than the prediction of just one person. Hence, an ideal user-contributed categorisation system will utilise what Surowiecki [132] calls *the wisdom of crowds*. That is, under certain circumstances, taking individual predictions made by a large group of people and aggregating them can result in a very accurate estimate.

For example, if we take a pair-wise mean of each player's DoM votes, and calculate and correlate these with the adjusted mean value* $\hat{d}$, then correlation jumps from 0.65 (GI) and 0.64 (NGI), to 0.77 and 0.76 respectively.

4.4.2 *Limitations of the Study*

As an exploratory study, this experiment showed some interesting results. There is much room for further expansion however:

- The experiment was only conducted with first-year engineering students from a single university. Females in particular, were under-represented in the group. A further study with a broader demographic range would be useful to address this.
- Only a relatively small number of artifacts were categorised due to time constraints. A study with a larger number of artifacts may allow more conclusions to be drawn regarding the effect of different media types on categorisation.
- It is also very difficult (perhaps even impossible) to separate comprehension time from categorisation. A study which uses only very simple artifacts for categorisation (such as pictorial symbols) may give more accurate timing information as the time taken to comprehend the artefact is minimised.
- Using the Dewey Decimal top-level categories may not have been the most suitable category system to choose, as the Dewey Decimal System was primarily devised as a classification system for shelving books. Developing a number of categories more suited to the artifacts to be categorised may provide clearer results.

* This adjusted mean excludes both players

4.5 CONCLUSION

In this study, I asked participants to categorise a number of information artifacts using one of two interfaces. One interface allowed only binary membership distinctions, that is, either 'yes' or 'no' (NGI), while the other allowed for graded degrees of membership (GI). The results showed that the GI provided more accurate results, however participants using both interfaces were not particularly good at predicting others' categorisations. The NGI proved faster to use than the GI, which may be important in time pressured situations.

When choosing an interface, much, of course, depends on the information retrieval methods that the categorisations are being provided for. If the artefact being categorised must reside in one physical location (like a book on a shelf), then multiple categorisations as used in this system may not be suitable. If, however, the retrieval system uses a search algorithm that calculates relevance scores, then graded, multiple categorisations may be extremely useful.

When it comes to overall system architecture however, it is interesting to note the low levels of correlation between pairs of players (0.43 for the GI and 0.41 for the NGI). These low values suggest that the categorisation predictions of a single individual are not sufficient to give a good representation of the whole population of users. Thus, I propose that aggregating categorisation scores from multiple users may provide much better results than relying on the judgement of a single individual. This theme is taken up later in the thesis when we examine social navigation and folksonomies.

5

CONTEXTUAL REASONS FOR CATEGORISATION PROBLEMS

What causes categorisation problems? In Chapter 3 we examined cognitive reasons for categorisation problems, noting that there is often a mismatch between cognitive categories and the concrete category structures we create in information systems. Cognitive mismatch does not account for the full range of problems we observed in Chapter 2 however. There, we saw that categorisation issues also arise from the the social, political, and cultural environment in which the information system operates.

In this chapter we begin by examining why people categorise at all. People categorise for a variety of different reasons, and these reasons largely determine the structure of the category scheme along with the way people interact with it. Hence, understanding reasons for categorisation is important in understanding why categorisation is so problematic.

Having examined reasons for categorisation, we then move on to examine the wide range of contextual factors that impact categorisation. In doing so we tie together observations from Chapters 2–4, along with illustrations from the literature. Understanding why categorisation problems occur then allows us to examine how to address these issues.

The designer of an information system needs to be aware of contextual issues when developing a category scheme. As we shall see, these issues are not new or unknown—they are well documented in the LIS literature and elsewhere. However, the LIS literature tends to make a number of assumptions about information systems that do not always hold. Many information systems (like SIMPRESS) do not have a dedicated expert available to categorise new entries and maintain the category scheme. Systems such as SIMPRESS also feature relatively small numbers of records and incorporate multimedia data such as digital photographs and CAD drawings. The literature, however, tends to focus on large systems featuring mainly text-based data, and assumes the presence of a dedicated expert. Based on this apparent gap in the literature, we formulate a problem definition for the thesis.

5.1 WHY CATEGORISE?

Categorisation is a means, not an end in itself. We do not categorise simply for the sake of having categories; we categorise for a purpose. We see this both at the cognitive level, and in the concrete categorisation schemes we interact with every day. Whenever we sort our washing, fill out forms, file our taxation receipts or shelve our books, we are categorising for a purpose. It is important to understand these purposes to understand why categorisation problems occur.

In this section we draw a distinction between cognitive categories, categories in the mind, and concrete categories which have some physical embodiment. Cognitive and concrete categories are strongly related, but serve differing purposes.

5.1.1 *Cognitive Categories*

At the cognitive level, categories are used for a variety of functions [84]. Categorisation is a basic tool used to facilitate just about everything people do with their brains. Hence, Lakoff can write: 'There is nothing more basic than categorization to our thought, perception, and speech' [70, p. 5]. In this section we examine cognitive purposes for categorisation as well as reasons for creating concrete categorisation schemes.

One critical function of categorisation is to recognise objects. Recognising an object allows us to determine how we might interact with it. For example, how I interact with a puppy differs from how I would interact with a bear cub. This is particularly important when we encounter new objects. If I encounter a small, white, pleasant-smelling object in a hotel room, then recognising it as a bar of soap allows me to use it for washing. It also prevents me from attempting to eat it. Recognition is perhaps the most basic of categorisation functions.

Another function of categorisation is prediction. People use what they know about an existing category to make inferences about a new member. For example, if the reference books I use generally have a subject index at the back, then when I encounter a new book I will look in the back for a subject index. Population stereotypes are a good example of this kind of prediction. People will make predictions about the behaviour of a new person they meet based on social categories such as race and occupation [119].

Categories also play an important part in communication [81], as discussed in Chapter 3. Any item one might wish to describe can be

given a variety of names, depending on level of abstraction and the purpose of the communication [21]. A tree might be a eucalyptus, just a tree, or even vegetation. Each of these labels implies a different level of categorisation, and communicates different information about the tree.

This kind of cognitive categorisation is mostly unconscious. People recognise objects, predict behaviour and choose words with little conscious effort. Only in problematic cases does this kind of categorisation come to our attention. However, we interact with concrete categorisation schemes much more consciously. When we create file folders in computer systems or organise filing cabinets or sort papers into piles, we are making conscious, deliberate categorisation decisions.

5.1.2 *Concrete Categories*

The reasons we create observable, concrete category structures are not fully explained by cognitive reasons for categorisation. Although the underlying cognitive functions are much the same, we create concrete categorisation schemes for a much broader range of purposes. Understanding these purposes helps us to understand why categorisation problems occur.

It is purpose that will shape the categorisation scheme, much more than the characteristics of items to be categorised. Any category system is simply one view, or one way of structuring some knowledge about a group of items [79]. The purpose of the categorisation scheme will determine which features are important (and which are unimportant) for distinguishing between groups. Different reasons for categorisation will result in different categorisation schemes.

Here, we outline five common reasons for creating categorisation schemes. This not an exhaustive list, but is indicative of the variety of purposes categorisation schemes serve. So, why do we categorise?

1. TO GAIN UNDERSTANDING. One reason for categorising is to organise knowledge about a subject. 'Recognition of resemblance across entities and the subsequent aggregation of like entities into categories lead the individual to discover order in a complex environment' [62, p. 518]. That is, grouping things together allows an individual to make generalisations about the group as a whole. These generalisations then allow distinctions between groups to be explored, comparing similarities and differences.

Often people will make these groupings concrete in tables, lists or some other kind of category scheme.

The structure of the category scheme encodes knowledge about the category members. Hierarchical classification schemes, for example, encode information in superordinate/subordinate relationships. Biological taxonomies are an application of this; created (and used) by scientists to understand plants and animals. Other schemes allow different kinds of relationships between categories, encoding different kinds of information. Hierarchies, trees, paradigms, and faceted analysis all imply different kinds of relationships between members *

2. TO ENABLE COMMUNICATION. We have already discussed the relationship between categorisation and language in Chapter 3; in addition to this, however, concrete categorisation schemes can be used to facilitate communication. People may use a categorisation scheme to negotiate a common vocabulary or nomenclature about a subject. Professional thesauri and ontologies are examples of this kind of categorisation, forming an agreement about how a group of people will use certain words.

   In this way a categorisation system is what Star [127] describes as a *boundary object*, that is, a concrete object that people with disparate world views have in common. This commonality gives people a reference point from which to engage in discussion. In doing so, they are able to negotiate shared understanding, making meaningful communication possible.

   Categorisation schemes have been identified as boundary objects by a number of authors [cf. eg. 4, 15]

3. TO ALLOW NUMERICAL ANALYSIS. Counting anything requires some form of categorisation. To count apples, for example, I must first distinguish apples from other things such as leaves or oranges. The things people count range from the very simple (such as apples) to the very complex. Bowker and Star [16], for example, describe the enormous complexity associated with construction of the International Classification of Diseases (ICD), used to count causes of death around the world. Whether simple or complicated however, counting (and thus categorisation) forms the basis of all numerical analysis.

* See Kwasnik [67] for a detailed discussion.

4. TO EXPLICATE PROCEDURES. In order to specify what procedures to follow in certain situations, it is necessary to classify the situations themselves [eg. 15]. For example, some organisations produce emergency procedure booklets that specify what to do in case of fire, chemical spills, armed holdup, bomb threats, etc. For each class of emergency, the booklet will give instructions on how to respond. Emergency procedures are a simple example; however, this kind of classification forms the basis for sophisticated tools and techniques such as workflow management systems [44, 138] and Case-Based Reasoning (CBR) [1, 30].

5. FOR ORGANISATION AND RETRIEVAL. This is where items are categorised in order to find them again later. We find this kind of categorisation not only in electronic systems, but also in tool-sheds, kitchens, libraries and filing cabinets. Categorisation keeps things tidy, and allows people to readily find things they need.

There are many different reasons for creating concrete category schemes. In the context of this thesis, categorisation for organisation and retrieval is our primary focus. However, it is important to understand other purposes for categorisation as well. A single categorisation scheme will often serve multiple purposes. Understanding these purposes is essential to understanding the reasons for categorisation problems.

5.2 CAUSES OF CATEGORISATION PROBLEMS

Categorisation problems often arise from the context in which categorisation is carried out. Categorisation schemes are created by people, people who exist in complicated social, political and cultural environments. Because of this complexity, the causes of categorisation problems intermingle and overlap. In this section we survey some of these overlapping contextual causes, using examples from the SIMPRESS system described in Chapter 2.

5.2.1 *Conflicting Requirements*

'Classification systems in general inherit contradictory motives in the circumstances of their creation' [16, p. 66]. A single categorisation scheme may have a number of different stakeholders, each with different interests in the categorisation scheme. Some stakeholders may

wish to retrieve information, while others wish to create reports and statistics. Even among the stakeholders creating reports and statistics, different people may be interested in different levels of detail. The design of a categorisation scheme becomes a compromise between all these divergent needs.

Bowker and Star [16] observed these issues when studying the construction of the ICD. For example, clinical specialists wished to know the breakdown of each disease strain, and hence required very specific categories, whereas public health statisticians wanted 'broader, action-oriented categories like nutritional deficiencies, environmental factors that could be changed, and so on' [16, p. 145]. Each group had different interests in the ICD, and pushed for a category structure to suit its particular purposes.

We can also see conflicting requirements in SIMPRESS. From the design engineer's perspective, the *concern type* characterisation was somewhat irrelevant. Since reviews of FMIs are made on a part-by-part basis, there is rarely any need for engineers to look for similar problems across different parts. For the novice operator however, the ability to look for similar problems may be extremely useful in the problem solving process. For the operator who simply wishes to enter the data as quickly as possible, the *concern type* is simply another delay in getting back to their 'real' work. Different stakeholders in SIMPRESS have differing perspectives on what is important.

The different perspectives of stakeholder groups affect both the design of the category scheme, and how people interact with it. Because the design of the category scheme is a compromise, it will never suit the purposes of any stakeholder group perfectly.

5.2.2 *Political & Social Consequences*

Categorisations have political and social consequences [16, 59, 131]. How people categorise their expenditures has consequences for taxation, and building zones create categorisations within cities, which have consequences for the environment and the people inhabiting those areas. When the consequences of categorisation become personally significant, achieving a desired outcome will take priority over accuracy.

The *concern type* categorisation in SIMPRESS did not produce many significant political or social consequences. A related SIMPRESS module however, recorded causes for machine downtime, creating consequences for the operators entering data. Recording reasons for stop-

pages allows reports to be created showing where time is being lost in the production process. However, when reports on downtime are made visible throughout the plant, using categories such as *operators on break* or *waiting for materials* can have serious consequences.

Bowker and Star [16] also observed similar issues when doctors recorded cause of death for a death certificate. In many cultures, recording suicide or HIV as a cause of death can result in significant distress for families. Hence, doctors would attribute death to vague, non-specific causes. Eventually this lead to the development of a double system, with doctors able to write one cause on the death certificate, and another elsewhere for statistical purposes.

The political and social consequences of categorisation have a significant impact on how people categorise.

5.2.3 *Perceptions of the System*

How people perceive a system directly affects how they interact with it. '[T]he attitudes of the organization and its employees towards the introduction of an IT system, and the way work is monitored, can affect whether a system is accepted and used to carry out the work' [78]. If I suspect that data I enter will be used to evaluate me negatively, then I will be reluctant to use that system. My perceptions govern my actions.

This was shown clearly in Chapter 2; the employees at the stamping plant were extremely busy, focussed on what they understood as their core tasks, and since they perceived SIMPRESS as a support tool (not directly contributing to production), entering data into SIMPRESS was given a low priority. Hence, engineers often delegated the task of data entry to administrative staff. This resulted in a situation where the people entering data into the system were not the ones with expert knowledge of the problem. Thus, the quality of categorisation suffered.

Bowker and Star [16] also relate a similar story: 'The task of filling in the death certificates ordinarily falls on the doctor who does not necessarily see the value in filling in a complex form to the degree of accuracy required. After all, this patient is dead; is the time not better spent on the living?' [16, p. 144]. The ICD statistics were not really relevant to the doctor, whose concerns lie with the patients. Hence, there is little incentive for giving careful attention to accurate completion of death certificates.

It is not only peoples' attitudes towards the system that affect cat-

egorisation, but also their perceptions of how it works. For example, when an operator raises an FMI in SIMPRESS, it must be assigned to somebody for action. The operator who raises the FMI however, does not decide to whom it is assigned. Suppose this operator imagines that SIMPRESS automatically determines who is assigned the FMI based on the *concern type*. This will impact how he categorises.

To illustrate, suppose our operator raises an FMI related to weld quality in subassemblies, a problem that is the result of a design flaw. The operator thinks that the FMI should be assigned to a design engineer. He expects however, that if he categorises the FMI under *weld quality* it will be assigned to the sub-assembly manager, whereas if he categorises the FMI under *design*, it will be assigned to a design engineer. The operator will categorise the FMI as *design*, since this will produce the desired outcome.

On the other hand, suppose, for example, a male operator imagines that the FMI is sent to a senior manager, who then assigns the FMI to the appropriate department.† The operator may then decide that the manager is clever enough to send the FMI to the right place. On the other hand, he may also decide that the senior manager does not know enough about the problem, skip the FMI process entirely, and talk to the engineers himself. The operator's perceptions of the system, both how it works and what it is for, determine how he interacts with it.

5.2.4 *Interpretation and Subjectivity*

A related issue to individual perceptions is the issue of subjectivity. Each person possesses a unique personal history. People are shaped by personal experiences that no-one else shares in quite the same way. Our personal history, our emotional state, and our physical health all impact on how we think and reason. Thus, when it comes to making interpretive judgements (like categorisation), there is always the potential for 'vagueness, misunderstanding, and the erroneous or suboptimal transfer of intended meaning' [2]. That is, the subjective nature of personal experience impacts how people categorise.

The degree to which individuals differ in opinion was demonstrated clearly in Chapter 4. Of the ten artifacts categorised by participants, there was no single artefact that received a unanimous positive categorisation. The correlation between an individual and any other in-

† This is what actually happens in SIMPRESS.

dividual's categorisations was only around 0.4. Even when viewing the same artefact in much the same setting people's categorisation choices varied significantly. The issue of subjectivity means that it is very difficult to predict how an individual will categorise something.

In addition to natural variations between people, the problem of subjectivity is further exacerbated by the issue of change.

5.2.5 *Environmental Dynamics*

In Chapter 3, we saw that our brains continually update the categories we create so that they better suit our purposes and accommodate newly acquired knowledge. Not only do our internal category structures change however, but the environment in which a categorisation scheme operates also changes. This is especially true at the organisational level. People leave the organisation, others begin employment; new product lines are released, while others are phased out; organisations change structure, from sole-operator, to partnerships, to publicly listed companies; they are are bought, sold and merged with other organisations. Each of these factors will change the organisational environment and impact categorisation.

Problems occur when the historical categories are no longer appropriate to the current situation. When this happens, people 'subvert the formal schemes with informal work-arounds' [16]. That is, people will develop localised strategies for dealing with problematic cases. These localised strategies will, naturally, be tailored to local needs, and may be at odds with the requirements of other stakeholders.

People use a variety of strategies to work around inappropriate category structures. A common informal work-around is the use of the *Other* category as we saw in Chapter 2. Another is to use an obsolete category as a substitute for creating a new category. Sometimes the categorisation will simply be ignored. If the category scheme does not change to suit the needs of people using it, then the people using it will change how the scheme is used.

5.2.6 *Prediction*

A related issue to environmental dynamics is prediction. Unless all the items to be categorised are known beforehand, creating a category system always involves predicting the future. Mai [79] points out that this is a key difference between biological taxonomies for

classifying plants and animal, and categorising information: '[T]he objects included in scientific classifications are, more or less, available when the classification scheme is constructed. When an object that is not included in the scheme appears, the object first goes through a definitional process in the sciences and is then included in the classification scheme on the outcome of the scientific discourse.' [79, p. 43] By contrast, when categorisation schemes are created for information systems, then only a relative few items are available. The designer must predict what kinds of items will be categorised in future and design the category scheme accordingly.

This issue was discussed in Chapter 2 with uneven category usage for the *concern type* categorisation. Nineteen of the 53 categories had not been used at all in the three-year period, and thirteen categories had been used only once. At the same time the categories *Other* and *Design* accounted for more than 80% of the categorisations. While some of the *Other* categorisations were attributable to poor training, the over-use of the *Design* category suggests unanticipated use of the system. Investigating this revealed that after a successful trial, SIMPRESS was made available to another department in the plant for which it was not originally designed. This resulted in increased use of the *Design* category. Hence, categorisation problems occurred because of a failure to predict future usage. This may not have been at all foreseeable for the system designers, yet the cause is still prediction failure.

5.2.7 *The Tedium of Data Entry*

Data entry may often be boring and tiresome, and bored, tired workers tend to make serious mistakes and 'cut corners' [102]. Thus, the quality of categorisation suffers. If the volume of data is large, then data entry quickly becomes monotonous and repetitive. There is very little to give one a sense of satisfaction in the work [53], and data entry work is often perceived as a menial, unskilled task [128].

Tedium is also related to people's perceptions of the system. If the task is perceived as important and valuable, then data entry may become less boring. This is especially the case if there is a strong personal incentive for entering the data. However, as seen earlier, data entry is often delegated to administrative staff, and the people entering the data are not the same as the people using it. Thus, there is little incentive to perform the job well, and it quickly becomes a tedious imposition.

5.2.8 *Summary*

Surveying the various contextual problems associated with categorisation is a somewhat bleak activity. We see that 'classifications are never innocent but streaked with arbitrariness and motivated by preconceptions and prejudices. Besides, they are constantly shifting, whether by design or in spite of our efforts to capture them' [93 quoted in 79]. People do not behave like ideal, rational actors, but attempt to optimise their position in whatever system they find themselves. Hence, the design of any real world categorisation systems will involve ambiguity and compromise.

5.3 IMPLICATIONS FOR INFORMATION SYSTEM DESIGN

Having surveyed a range of causes for categorisation problems, a natural question to ask is 'What can we do about them?' Of course, a lot depends on who is meant by 'we': for the purposes of this thesis, 'we' are those involved in designing information systems. This may be software engineers, information scientists or librarians.

To clarify, any categorisation scheme will have at least three distinct stakeholder groups:

1. The designers who develop the category scheme (sometimes called 'classificationists');
2. The people responsible for categorising new entries into the system (sometimes called 'classifiers'); and
3. The people who use the system.

Sometimes two of these groups will be combined. Classificationists may also be classifiers, and classifiers may also be users. It is rare for the users to be classificationists however. This has the potential for creating difficulties, since the classificationist will often hold an entirely different world-view from the users. They will have different skills, expertise and language. Thus, to create a categorisation scheme, the designer must first seek a deep understanding of the domain: 'To create a classification system for a particular company, organization, library, or any other information center, one needs to begin with a study of the discourse and the activities that take place in the organization or domain' [79]. Such a study allows the designer to gain an understanding of the contextual factors that can potentially cause problems.

In reality however, the system designer in this situation has only a limited sphere of control. Effectively, we assume that they exercise control chiefly over the design of the system. Hence, it is beyond the designer's scope to change organisational culture, or to change the conflicting requirements of stakeholders. However, the factors described in Section 5.2 must be at least understood when designing any information system.

What then *can* the system designer control? Here we identify five aspects of an information system that affect categorisation quality:

CATEGORY STRUCTURE. Category structure refers primarily to the relationships that are possible between categories. A hierarchical classification scheme, for example, requires categories to be mutually exclusive and creates a hierarchical ordering. An uncontrolled vocabulary, on the other hand, allows multiple categorisation and specifies very few relationships between categories. So, questions relating to category structure include:

- Will categories be mutually-exclusive, or will multiple category membership be possible?
- Will category membership be binary (in/out), or fuzzy (graded)?
- Will categories be arranged hierarchically or will all categories be at the same level of abstraction?
- Which users (if any) will have the ability to create new categories?

Each of these questions will need to be answered with reference to the requirements of stakeholders and the nature of the items for categorisation.

USER INTERFACE. The user interface provides the means for users to make categorisations. To a large extent, this is constrained by the category structure. However, there are still many options even once structure has been determined. Radio boxes and drop-down lists, for example, both provide for mutually-exclusive categorisations and are useful in different situations. In Chapter 4 we saw that the categorisation interface affects both the speed of data entry and the degree of consensus. Hence, careful attention must be paid to the design of the user interface.

VOCABULARY. Vocabulary refers to the labels given to categories. Vocabulary is important since the vocabulary and structure define

what can be said about the items being categorised. This implicitly defines what *cannot* be said about an item. Hence 'to suggest a classification of a knowledge field [...] is—in one way or the other—to support given theoretical viewpoints at the expense of other views' [59].

Depending on what category structure is chosen, the designer may abdicate some level of control to the users. Where users are able to freely use any labels they choose, this is referred to as an uncontrolled vocabulary. Conversely, a scheme where the designer specifies all the category labels is called a controlled vocabulary [133].

Uncontrolled vocabularies have the advantage of closely reflecting users' vocabulary. There is no problem of the designer imposing unfamiliar labels through the information system. At the same time, however, studies have shown that people tend to produce a wide variety of labels for the same object [eg. 27, 42]. This can lead to messy categorisations and poor retrieval.

On the other hand, a controlled vocabulary has its own challenges. There is always the potential for mistranslation between the designer's vocabulary and the user's [22]. To combat this, a system designer needs to understand the user's world in some detail. '[The] construction of classification schemes needs to rest on studies of users' information interactions, work and habits, as well as, the structures of domains' [79, p. 46].

GRANULARITY. Granularity refers to the level of detail that categories provide. As seen in Section 5.2.1, finding an appropriate level of granularity will often require a compromise between various stakeholders. A coarse granularity usually results in less categories and can make categorisation quicker for those entering data. However, a coarse granularity may obscure important information and render the categorisation less valuable.

FLEXIBILITY. Flexibility is the ability of the categorisation scheme to respond to change. As discussed in Chapter 3, all categorisation schemes require mechanisms for dealing with category shifts and changes. This will vary depending on the organisational environment and the items being categorised. However, all categorisation schemes will change at some point and some mechanism must be available to facilitate it.

There are a number of different approaches to facilitating change:

- The system designer can carry out periodic reviews and analysis of the system and reconstruct the categorisation scheme to match current needs.
- The organisation can entrust a few administrators with the task of adding and removing categories from the system as necessary.
- Using an uncontrolled vocabulary removes the need for anyone to update categories since users may create a new category whenever they wish.

The choice of approach will constrain what kinds of change are possible, and the level of effort required to maintain the system.

While a system designer may only have limited control over the context of an information system, they have a much greater degree of control over the system design. The design of categorisation schemes must be carefully matched to the requirements of the organisational setting.

5.4 PROBLEM DEFINITION

Having surveyed a range of categorisation problems at the cognitive and contextual levels, we now set out the aim of this thesis. The categorisation problem is by no means new or unknown. It is well documented in the cognitive science and LIS literature, as we have seen. However, the LIS literature in particular tends to make a number of assumptions:

- That a dedicated librarian or administrator is available to maintain and modify the categorisation structure. Given that library and information science professionals are in the business of creating and maintaining classification schemes, it is not surprising that the literature makes this assumption. In many cases, the literature assumes that a dedicated expert categorises all new entries into the system, thus avoiding many of the issues discussed in Section 5.2.
- That the items to be categorised are primarily textual documents. Thus, the literature has a lot to say about interpreting, understanding and indexing documents. However, as we saw with SIMPRESS, many information systems are now incorporating more and more multimedia data.

- That the number of items to be categorised is very large. Managing large sets of documents presents unique and difficult challenges. However, as seen with SIMPRESS problems still occur in smaller-scale systems. Furthermore, the solutions applied to large systems do not always work well with small ones.

In light of these assumptions, we propose that the scholarly literature has largely neglected smaller-scale systems without a dedicated librarian. Hence, the aim is to develop methods for categorising in smaller scale information systems that meet the following requirements:

1. *Reflects user vocabulary.* Most categorisation systems attempt to reflect user vocabulary, with varying degrees of success. As seen in Chapter 4 however, individual vocabulary varies so greatly that reflecting user vocabulary will always be a difficult task.

2. *Does not require a dedicated librarian/administrator to maintain and modify the category structure.* Having a dedicated librarian is not always feasible for smaller-scale categorisation schemes—particularly in the case of SMEs. In these circumstances, the people performing categorisation tasks are the day-to-day users of the system.

3. *Responds to changes in vocabulary or the kinds of items categorised.* If a librarian or administrator is not available then the categorisation method must be able to accommodate different kinds of items as the environment in which the system operates changes over time. It must also be able to respond to corresponding changes in vocabulary.

4. *Is suitable for multi-media data.* Information systems are incorporating more and more multimedia data. For example, SIMPRESS relied heavily on digital photographs and scanned blueprints (among other things). Hence, the categorisation method must be suitable for non-textual data.

5. *Is suitable for small numbers of records.* It is sometimes assumed that where there is only a 'small' number of items, organisation becomes trivial. As seen in Chapter 2 however, this is not always the case. Thus, the method should be able to work effectively with as few as 200 records.

In the next chapter we survey approaches in the literature which may address these requirements.

6

POTENTIAL SOLUTIONS

What kind of categorisation scheme will meet the requirements set out in Chapter 5? This chapter surveys various approaches found in the literature from a range of different research communities. After weighing the advantages and disadvantages of each approach, *folksonomies* appear to be the best candidate for addressing the aims of this thesis.

In the last chapter we formulated our problem definition, that is, the aim of this thesis is to find categorisation methods that:

A. reflect user vocabulary;

B. are able to adapt to change;

C. do not require expert human intervention;

D. are suitable for multimedia data; and

E. work effectively with small numbers of records.

This chapter weighs the advantages and disadvantages of various approaches to categorisation, building on work from LIS, HCI, machine learning, and information architecture. Our survey of the literature shows that there is no current approach that completely fulfils all these requirements; there are no silver bullets with information systems. However, the folksonomy approach shows the most potential for addressing the thesis requirements.

6.1 ECOLOGICAL CLASSIFICATION SCHEMES

Ecological Classification Schemes are schemes which use the context* of the information system as the basis for creating a classification scheme. Using this kind of approach, the system designer conducts an in-depth study of the context in which the information system operates. 'The study and analysis needs to go beyond just the activities that involve the classification scheme to understand the work domain

* I use the term 'context' here to be consistent with other chapters of this thesis. Authors in the literature use various terms to describe similar concepts such as 'domain' [58], 'discourse community' [79], 'ecologies' [134], and 'work context' [39, 40].

in full' [79]. The result is a classification scheme which closely reflects the vocabulary and work habits of people using the system.

The phrase 'Ecological classification schemes' actually refers to a group of analysis methods including Domain Analysis [55–58] and Ecological Work Analysis [3, 5, 6, 39, 40]. In labelling them Ecological Classification Schemes, we follow the notation of Mai [79] and Tennis [134]. The common theme in these approaches is that the unit of analysis is the *context* or *ecology* (at various levels), rather than just users or documents. These approaches also stress that the classification scheme should reflect the work habits of people who use the system, rather than being imposed by an epistemic authority.

Domain Analysis was introduced by Hjørland and Albrechtsen [58] in an attempt to make explicit some of the changes to approaches and assumptions occurring in LIS. This mainly constituted a shift away from positivistic (or rationalistic) theories of knowledge towards a relativistic or post-modern epistemology [58, 79]. Under positivistic theories of knowledge, it was assumed that scientific theories could be discovered that would produce an optimal classification scheme solely by examining the nature of the items to be classified. Relativistic theories of knowledge however, assert that knowledge is inseparable from language and context [20, 144]. Reflecting this shift, the focus of classification research broadened from looking solely at documents and individuals to include the larger discourse community.

The focus of domain analysis is to understand the activities and language of a particular knowledge domain:

> 'The domain-analytic paradigm in information science (IS) states that the best way to understand information in IS is to study the knowledge-domains as thought or discourse communities, which are parts of society's division of labor. Knowledge organization, structure, cooperation patterns, language and communication forms, information systems, and relevance criteria are reflections of the objects of the work of these communities and of their role in society. The individual person's psychology, knowledge, information needs, and subjective relevance criteria should be seen in this perspective.' [58, p. 400]

Hjørland and Albrechtsen [58] contrast this focus on the knowledge-domain with 'cognitivism', an approach that focuses solely on the psychology of the individual to the exclusion of the 'disciplinary context'. Cognitivism attempts to develop models of a 'typical' or 'ideal'

user, and to design information systems to suit these models. However, as seen in Chapter 4, individuals vary so greatly as to make this kind of modelling impractical. At the broader level of the domain however, groups of people tend to agree *in general* on vocabulary and structures of knowledge. By focusing on discourse communities rather than individuals, domain analysis allows development of a classification scheme that suits the community, rather than a non-existent ideal individual.

While domain analysis is a powerful methodology, it focuses mainly on classification of scientific publications [cf. e. g. 57], and is not really suited to enterprise contexts. In response to this, Albrechtsen and colleagues [3, 5, 6, 104] proposed Cognitive Work Analysis (CWA) as a framework for studying the context of an enterprise or work group.

According to Albrechtsen and Pejtersen [5], p. 215, CWA is 'a methodology for systematic exploration and analysis of work domains'. That is, CWA is like a set of tools for analysing a work situation, giving a holistic picture of a work domain: What gets done, who does it, how they do it, and why. This analysis then informs the design of an information system to suit the particular domain.

CWA is described as a holistic, multidisciplinary approach because it attempts to analyse many different aspects of a work domain *simultaneously* [39, 40]. These aspects include:

- The tasks people perform;
- The environment in which tasks are carried out;
- The people who carry out the tasks; and
- The reasons for their actions.

These different aspects are often presented using *the onion model* (see Figure 13). Each layer of the 'onion' becomes more specific, starting with the external environment and the organisation, moving through to perceptions and cognitive attributes of individual people at the centre. Fidel et al. [40], p. 943 provides a good summary of typical questions asked at each stage of the analysis (see Table 22).

Because work domains are complex, dynamic systems, they are often likened to ecologies, hence the phrase *ecological work based classification schemes* [104].

The advantage of these approaches (domain analysis and ecological work analysis) is that they produce a classification scheme well suited to the vocabulary and work habits of the people using it. Studying the work habits and vocabulary of people using the system gives

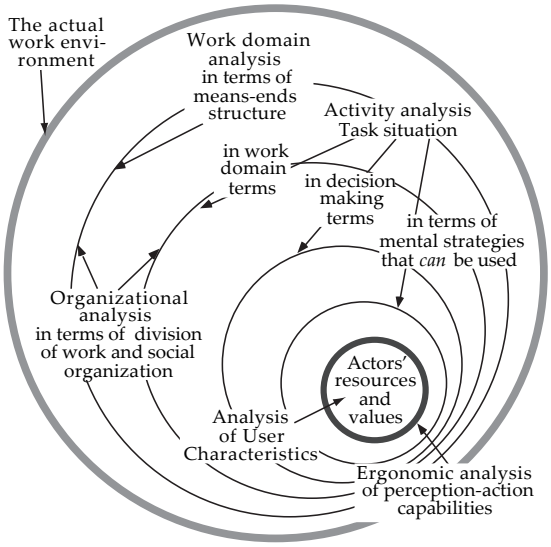


Figure 13: The multiple perspectives involved in Cognitive Work Analysis (CWA) (the Onion Model). Adapted from Pejtersen and Fidel [105].

Table 22: Example questions asked at each level of Cognitive Work Analysis (CWA). Adapted from Fidel et al. [40].

ASPECT	EXAMPLES OF QUESTIONS TO ASK IN ANALYSIS
Environment	What elements outside the organisation affect it?
Work domain	What are the goals of the work domain? The constraints? The priorities? The functions? What physical processes take place? What tools are employed?
Organisational analysis	How is work divided among teams? What criteria are used? What is the nature of the organisation: hierarchical, democratic, or chaotic? What are the organisational values?
Task analysis in work domain terms	What is the task (e. g., design of navigation functionality)? What are the goals of the task that generated an information problem? Constraints? The functions involved? The tools used?
Task analysis in decision-making terms	What decisions are made (e. g., what model to select for the navigation)? What information is required? What sources are useful?
Task analysis in terms of strategies that can be used	What strategies are possible (e. g., browsing, the analytical strategy)? What strategies does the actor prefer? What type of information is needed? What information sources does the actor prefer?
Actors' resources and values	What is the formal training of the actor? Area of expertise? Experience with the subject domain and the work domain? Personal priorities? Personal values?

designers an in-depth understanding of the community's knowledge structures. The classification scheme can then be designed to match these structures, rather than requiring people to adjust to a new classification system.

Are ecological classification schemes a solution to the categorisation problem then? While ecological approaches address some of the requirements identified in Chapter 5, they are more methodology than method. That is, ecological approaches tend to specify what kinds of questions to ask, rather than identifying solutions. While they provide an excellent framework for approaching information system design, methods, tools and techniques are still required. Ecological approaches simply lead us to ask the same questions proposed by this thesis: What categorisation schemes are best suited to situations where there is no dedicated administrator, lots of multimedia data, and a small number of users?

As pointed out by Tennis [134], ecological approaches also fail to address the flexibility requirement. An in-depth study of the discourse community uncovers the way people's knowledge is structured *now*, but says nothing about how knowledge will be structured in the future. It is entirely possible to create a categorisation scheme that will be obsolete as soon as it is implemented. An ecological analysis may discover the need for flexibility, but does not really suggest how to implement it.

As methodology, ecological approaches provide a good starting point for our investigation. However, they do not offer solutions so much as suggest where to look to define the problems. In order to find solutions we must look more specifically at categorisation methods and techniques.

6.2 FACETED ANALYSIS

An important approach to categorisation discussed in the literature is faceted analysis and Faceted Analytico-Synthetic Theory (FAST).

The idea of Facet Analysis was first proposed by S. R. Ranganathan in the 1930s [107], and published more fully in the 1967 edition of *Prolegomena to Classification* [108]. This work was further expanded and revised by the Classification Research Group (CRG), formed in 1952 [19, 41, 126]. The approach gained significant popularity in LIS, and more recently has been applied to website design and information architecture [68, 69, 99].

The formal rules and principles for facet analysis, published by

Ranganathan and the CRG, are somewhat complicated and difficult to read [126]. This has led a number of authors to write simplified or summarised versions of the analysis process [e. g. 31, 67, 126], each with their own modifications and variations. Thus, attempting to find a good introduction to the field can be a difficult process [69].

Denton [31] describes facet classification as 'a set of mutually exclusive and jointly exhaustive categories, each made by isolating one perspective on the items (a facet), that combine to completely describe all the objects in question...'. In other words, the set of items to be classified can be viewed from a number of different perspectives. In facet analysis, a separate classification scheme is created for each of these different perspectives. These separate classification schemes are called *facets*.

The advantage of facets in a digital environment is that they can be combined in any order to suit the individual. This is particularly useful in online stores. For example, Morville [99] uses the example of an online wine store to illustrate different facets for wine (see Table 23). Here, one customer may wish to browse wines by type, knowing that they wish to purchase a Pinot Noir, another customer may wish to browse wines by price, knowing that they have a budget of $30.00, and yet another customer may wish to look for a 2000 Australian Shiraz at any price. Having *orthogonal* facets allows these criteria to be combined in different ways.

Table 23: Facets for an online wine store. Adapted from [99]

FACET	SAMPLE VALUES
Type	Red (Merlot, Pinot Noir), White (Chablis, Chardonnay), Sparkling, Pink, Dessert
Region (origin)	Australia, California, France, Italy
Winery (manufacturer)	Blackstone, Clos du Bois, Cakebread
Year	1969, 1990, 1999, 2000, 2007
Price	$11.95, $31.95, <$200, Cheap, Moderate, Expensive

This simplified understanding of facets provides a useful tool for organising websites. In its original form however, FAST was originally intended as a comprehensive method for categorising any body of information. The goal was 'to create a a hierarchical ordering or mapping of all knowledge' [68], such that any information artefact could be placed in its 'correct' location. To facilitate this goal, Ranganathan proposed five universal facets for organising everything:

PERSONALITY: core concepts representing items.

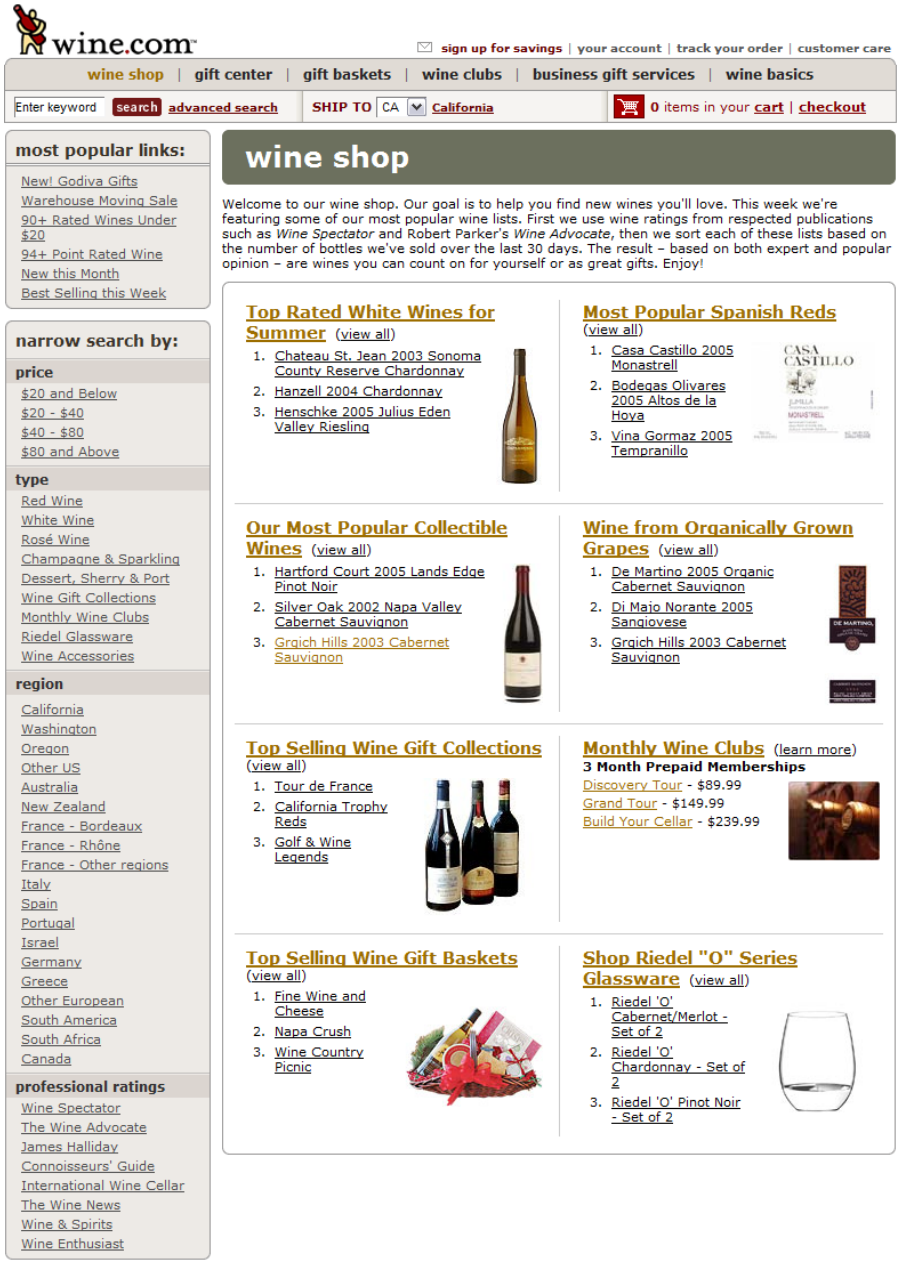


Figure 14: Screen capture from wine.com showing faceted navigation on the left panel.
(http://www.wine.com/wineshop/default.asp?hid=hpa_wineshop. Retrieved 12 June 2007.)

MATTER: substances or materials.

ENERGY: activities or actions.

SPACE: physical location.

TIME: dates etc.

The CRG later expanded these universal facets to thirteen: *thing, kind, part, property, material, process, operation, agent, patient, product, by-product, space, and time*. Each of these universal facets may have sub-facets. Under the *material* facet, socks, for example, may have different fabric (wool, polyester, cotton, etc.), different colours (black, gray, brown, blue), and different patterns (plain, striped, spotted, checkered). Following classical facet analysis, items should be listed (or displayed) according to the arrangement above: first by *thing*, then by kind, part, property, material, and so on. According to Broughton [19], p. 55, the three major theoretical arguments for this order are:

1. The order progresses from concrete to abstract;
2. Each facet is dependent on preceding facets; and
3. It gives the most useful grouping of compounds.

In most circumstances, however, restricting a classification scheme only to these universal facets is impractical. Not all of the thirteen facets will be useful, and in some cases more divisions will be necessary. The CRG recognised this, and the thirteen basic categories are generally considered a starting point for 'discovering' appropriate facets [31]. In most modern faceted classifications facets are chosen pragmatically to suit the domain.

The main requirement for a faceted classification is that facets be both *exhaustive* and *orthogonal*. Facets must be exhaustive so that they include all possible items to be classified; ideally, this means that there will be no *Other* categories. Facets must be orthogonal so that there is no overlap or interaction between facets. For example, listing a wine as 'red' does not determine what price it should be. These requirements are part of what makes facets so powerful. Orthogonality ensures that facets can be arranged in different orders to suit individual requirements, and exhaustiveness ensures that every item has a place in the classification scheme.

Based on these requirements, it is often claimed that facets are extremely flexible. In the sense of adjusting to individual requirements,

this is certainly true. In the sense of adjusting to changes in vocabulary and types of items however, flexibility is limited. The orthogonality requirement ensures that if a new class must be added to a facet to accommodate a new item, or items, then the change to this facet will not affect any other facets. Thus, classes can be added or removed without affecting the entire classification scheme—the changes are restricted to one facet only. In this sense, facets restrict the impact of change. However, some person must still decide that this change is necessary, leaving the question: Who will do it?

Coming from the LIS literature, facet analysis assumes that a trained professional is available to do the work of creating and maintaining the classification system. While it is relatively flexible in that the entire scheme does not need to be re-built from scratch every time a new class is required, it still relies on maintenance by a skilled administrator.

6.3 CARD SORTING

Card sorting is a method for gaining insight into users' mental structures and has become widely used in Information Architecture and HCI communities. Maurer and Warfel [87] define card sorting as follows:

> Card sorting is a user-centered design method for increasing a system's findability. The process involves sorting a series of cards, each labelled with a piece of content or functionality, into groups that make sense to users or participants.

Card sorting was originally developed by the Knowledge Engineering community as a tool for eliciting tacit knowledge from experts [cf. 90, 115, 145]. Researchers proposed the technique as a tool to help solve the 'knowledge acquisition bottleneck' involved in developing a knowledge based system [115]. Along with the method for conducting card-sorting sessions, researchers also suggested sophisticated analytical techniques for processing the resulting data.

The first group to use card sorting in a web context appears to have been Nielsen and Sano [100] when developing the structure for Sun Microsystem's intranet. Since then the technique has gained popularity in the information architecture community as a user-centred technique to help plan the structure of websites [87].

The steps in undertaking a card sort are as follows:

1. Select content to be sorted (eg. web pages, online manuals, library documents, etc).
2. Select participants (individuals or groups of people who will use the system).
3. Prepare the cards, writing out headings or summaries on cards small enough to be sorted easily.
4. Ask participants to sort the cards into groups and label their groups at the end.
5. Analyse the results.

Different authors recommend different approaches to card sorts. Some suggest getting individuals to do card sorts on their own, while others suggest that a group should do the sorting. Still others suggest asking multiple groups to perform card sorts.

The advantage of having individuals perform card sorts is that it is possible to get a numeric measurement of how much people agree or disagree. The disadvantage of having individuals perform card sorts is that it creates a lot of data whose analysis can be time consuming. Research by Tullis and Wood [136] suggests that between 15–30 participants are needed before a stable consensus is reached. Because this is a large amount of data to analyse, a group sort may be more cost effective.

Another advantage of group sorts is that the researcher can make observations as the group interacts. The group may argue about where certain items should be placed, and these items can be noted for future reference along with the reasons for the controversy. The researcher can also ask the group to explain its reasoning for various items or groups, giving further insight into the group's mental models.

Card sorts can also be performed to uncover differences between diverse groups of people, for example, Upchurch et al. [137] used card sorting methods to compare farmers perceptions of a website with those of web designers and 'lay people'. If different participant groups show significant variation, then card sorting may uncover a need for multiple categorisation schemes (or a faceted classification).

The advantage of card sorting is that it is simple to understand, easy to implement, and relatively inexpensive. It is also an extremely flexible approach, with many different variations. The results of card sorting can be analysed using statistical methods which often give the perception of being 'scientific' and therefore reliable.

Asking participants to label their groupings can be a useful way of learning user vocabulary. Using these labels to mark categories in the system can help prevent mismatch between the designer's vocabulary and the users'. This in turn helps to improve the findability of items in the information system.

One disadvantage of this approach is that it is a content-focussed technique, rather than a task-focussed technique [87]. That is, the card sorting method is focussed on what is *in* the system, rather than what people *do* with the system. This means that while a card sort will result in a logical structure for content, this structure may have no relation to how people use or access the content; the categorisation scheme may end up suiting the content rather than the people using the system.

Another disadvantage of this approach is that it really only provides useful information, not a categorisation scheme. It may be a helpful tool for assisting with ecological work analysis (as discussed in Section 6.1), but like those approaches it does not suggest what the best structure for the category scheme will be.

Also, like ecological approaches, card sorting provides a snapshot of users' current mental models. It does not show how those mental models will change, or suggest how a category scheme can adapt to those changes.

6.4 AUTOMATIC TEXT CLASSIFICATION

One of the recurring problems with the approaches described above is the need for a skilled administrator to look after the categorisation system. If this is such a prevalent issue, then it seems logical to suggest automatic classification as an appropriate solution. If nobody is available to do the work, then why not get a computer to do it? Automatic classification methods seek to place documents in a classification scheme based on an automated analysis of the document text.

There are a number of different methods for automatic text classification. Lubbes [77] differentiates between rule based systems and pattern matching systems. *Rule based systems* are similar to knowledge engineering approaches, where human experts define a set of rules for classifying documents. These were particularly popular until the late 1980's, but increasingly lost popularity in the 1990's in favour of pattern matching approaches [117]. In rule based systems, the computer parses the document to identify user-specified entities

(such as keywords or subject headings), and assigns the document to an appropriate category based on the rule set.

Pattern Matching Systems are based on machine learning and IR techniques. They use some method to transform each document into a mathematical representation, and match this representation to an appropriate category. There are a number of different ways of doing this, including probabilistic classifiers, symbolic classifiers, k nearest neighbours (k-NN) algorithms, neural networks, and support vector machines (SVMs).

6.4.1 *Rule Based Systems*

To create a rule based classifier a knowledge engineer or records manager works with a domain expert to develop a set of classification rules. The classifier then parses textual documents and classifies them according to this rule set. Rules are generally of the form:

IF ⟨*conditional statement*⟩ THEN ⟨*action*⟩

Usually the conditional statement refers to the presence of a particular phrase in the text, while the action is to classify the document in a particular category. Table 24 shows some example rules.

Table 24: Example rules for a rule based classifier

Rule	IF	Conditional Statement	THEN	Action Statement
1.	IF	'coffee' appears in the document	THEN	classify in *Coffee*
2.	IF	'programming' appears in the document	THEN	classify in *Programming*
3.	IF	'SDK' and 'java' appear in the document	THEN	classify in *Programming*
4.	IF	'java' and 'roast' appear in the document	THEN	classify in *Coffee*

One advantage of rule based approaches is that they are able to take advantage of document metadata. Rules can be specified which look at the document author, title, or keywords, as opposed to the entire text. These kinds of rules are common in email systems (e. g. 'IF subject-line contains "viagra" THEN classify in SPAM').

Another advantage of rule based approaches is that they are a transparent process. That is, anyone can examine the rule set to see why

a document has been classified in a particular class. With pattern matching approaches however, it is much more difficult to determine why a particular document has been placed in a particular class.

A major issue for rule based systems is that as the number of categories grows larger, then the rule set for a rule based classifier can become very complex. More specific conditions and larger rule sets are required to differentiate between classes. In a review of machine learning approaches to automatic classification, Sebastiani [117], p. 8 writes:

> The drawback of this approach is the *knowledge acquisition bottleneck* well known from the expert systems literature. That is, the rules must be manually defined by a knowledge engineer with the aid of a domain expert (in this case, an expert in the membership of documents in the chosen set of categories): if the set of categories is updated, then these two professionals must intervene again, and if the classifier is ported to a completely different domain (i. e., set of categories), a different domain expert needs to intervene and the work has to be repeated from scratch.

Because of these disadvantages, Sebastiani argues that pattern matching systems are a more useful approach.

6.4.2 *Pattern Matching Systems*

In a pattern matching system the classifier must first be *trained* to recognise the characteristics of each category. To do this, an information professional must provide the classifier with a set of example documents (called a *training set*) already classified by human intervention. The classifier then generates an internal representation for each class. Any new document is then compared to these internal representations and the classifier assigns it to an appropriate class.

The selection of a good training set is crucial to the effectiveness of the classifier. This requires an information professional to be familiar with both the corpus to be classified and the operation of the classifier. The different methods of pattern matching operate in quite different ways to generate their internal representation of a class.

The steps in training a classifier are illustrated in Figure 15. First a classification scheme must be created by a human expert to accommodate the documents. Some automated tools are available to assist with this, but significant human input is still required to create an

appropriate scheme [52, pp. 342–345].

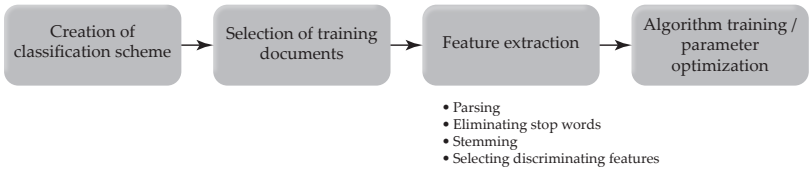


Figure 15: Steps in training an automatic classifier

Next, training documents must be selected to represent good examples (and possibly good non-examples) of each class. In some cases it may even be necessary to remove paragraphs or sections of documents to create training-only examples.

After this, the documents must be transformed into a mathematical representation. This process is called *feature extraction* [76] or *document indexing* [117], and involves:

- Parsing the document into individual linguistic components;
- removing stop-words (e. g. 'the', 'of', 'a', 'it');
- stemming (e. g. 'search', 'searching', 'searched', 'searches', 'searcher' → 'search');
- selecting good phrases or terms to distinguish between documents; and
- Representing the document as a feature vector.

A *feature vector* is something like a signature for the document. Each number in the vector represents a magnitude associated with a feature. For example, suppose the feature extraction process identified 'roasting method', 'Arabica beans', 'object oriented' and 'class inheritance' as highly discriminating features. Every document in the corpus would have an element in its feature vector for each of these phrases. This is illustrated in Tables 25 and 26.

Feature extraction may also include other refinements to improve the selection of appropriate features. For a more detailed discussion see Sebastiani [117], pp. 10–18.

The final stage of the training process is tuning the classifier. This involves running the learning algorithm on the training set and adjusting parameters until a good result is achieved. The specifics of how this is done depends on the classification method chosen.

Table 25: Discriminating Features

Article Title	Roasting Method	Arabica Beans	Object Oriented	Class Inheritance
Roasting Your Own Coffee	8	3	0	0
Zen and the Art of Espresso	2	1	0	0
OO Programming Using Java	0	0	4	10
Java and Java Beans	0	0	6	2

Table 26: Feature Vectors

Article title	Feature Vector
Roasting Your Own Coffee	[8, 3, 0, 0]
Zen and the Art of Espresso	[2, 1, 0, 0]
Introduction to OO Programming Using Java	[0, 0, 4, 10]
Java and Java Beans	[0, 0, 6, 2]

A detailed discussion of the various pattern matching approaches is beyond the scope of this thesis. See Sebastiani [117] for a good introduction. Here, we will give a brief overview of some popular approaches.

Rocchio Method

The Rocchio and k-NN algorithms are perhaps the easiest to understand because they can be explained geometrically. The Rocchio method in particular can be easily shown on a simple graph. This is in contrast to probabilistic classifiers or neural networks, which are more difficult to explain using a diagram.

After feature extraction, each document in the training set can be represented using a vector. If there were only two discriminatory terms, it might look something like Figure 16. For each class the learner calculates a centroid, and draws an N-dimensional hypersphere around it (where N is the length of the feature vector). When a new item is to be classified, the classifier simply extracts its features, then calculates whether or not the item falls within a class sphere. The item is then placed in the appropriate class, or labelled as miscellaneous.

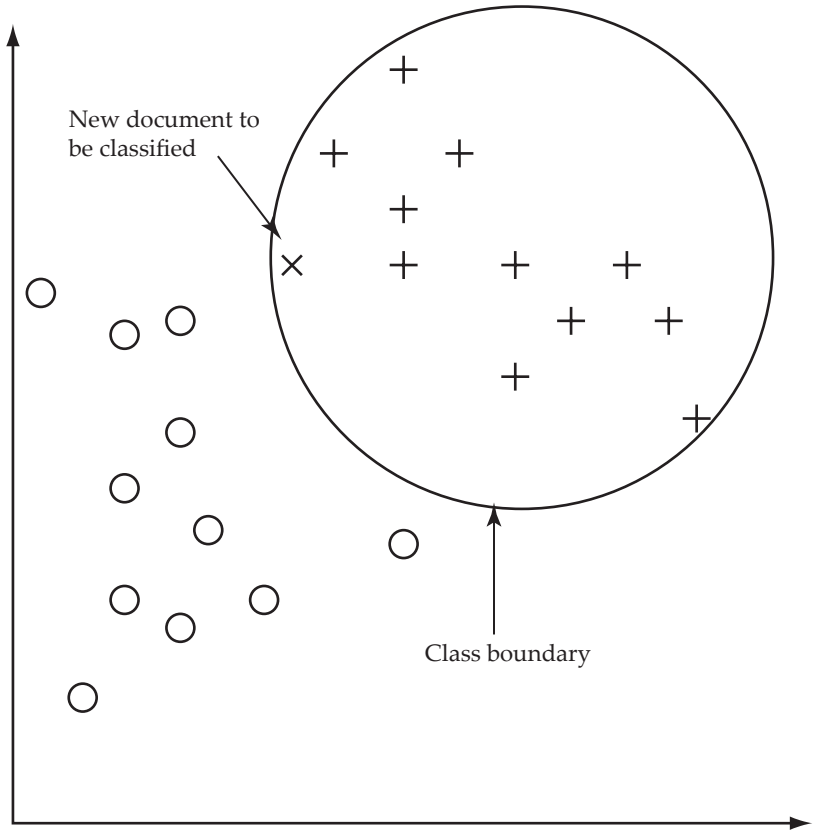


Figure 16: Example of the Rocchio method of automatic classification

k *Nearest Neighbours Algorithms*

The k *nearest neighbours* (k-NN) algorithm is similar to the Rocchio method, except that it does not draw spheres around the centroid of each class. Instead, when a new document is to be classified, it calculates the distance of that document from the k nearest training documents (Figure 17). If the k nearest training documents all belong to the same class, then it places the new document in that class too. If only some of the k nearest documents belong to another class, the classifier may use a threshold to decide whether or not to place it in the class or mark it as miscellaneous.

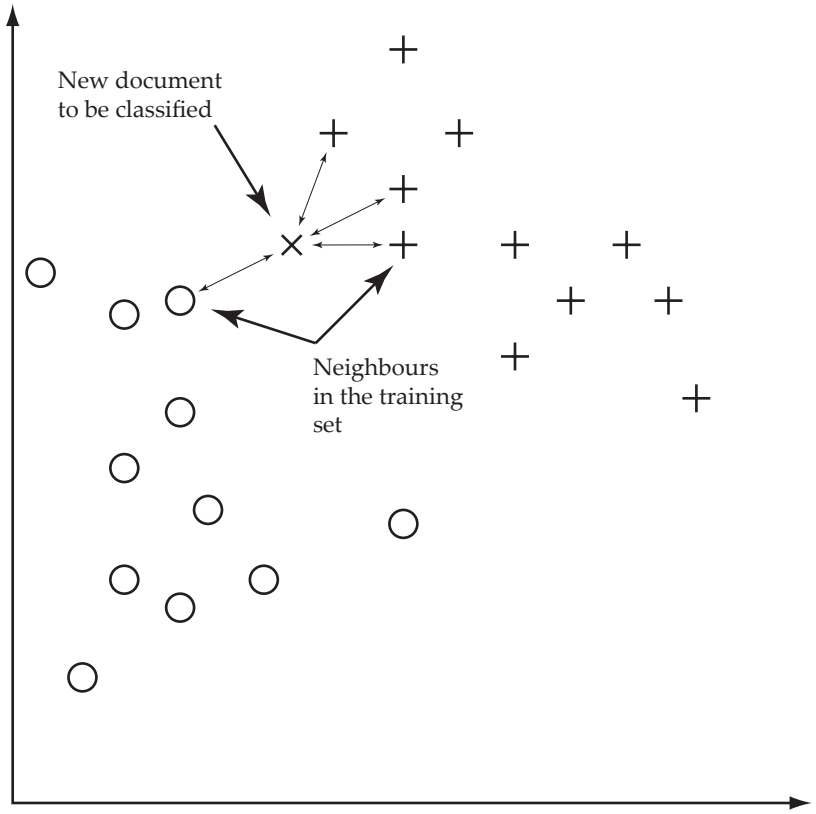


Figure 17: Example of the k nearest neighbours (k-NN) method of automatic classification

Probabilistic Classifiers

Probabilistic classifiers, also called *Bayesian classifiers*, work on the assumption that knowledge about the distribution of previous outcomes helps predict future outcomes. That is, knowledge from the past can help predict the future. In the context of automatic classification, a

probabilistic classifier builds a statistical model based on the training documents, which is used to predict which class new documents will belong to.

Usually the statistical model is created by assuming that each column in the feature vector is an independent variable. The learner then models the distribution of this variable, using these distributions to classify new documents.

One drawback to this approach is that terms in a document are generally not independent variables. Attempts have been made to create probabilistic classifiers which do not make this assumption, but they tend to be very complicated and computationally expensive.

Symbolic Classifiers

Symbolic classifiers attempt to automatically learn rules similar to those created in a rule-based system. There are two main types of symbolic classifiers: decision trees and inductive rule learners.

Using a decision tree, a document is classified by starting at the top node and moving down the tree based on what features are present in the document (see Figure 18). When the classifier reaches a leaf node, it places the document in the appropriate class.

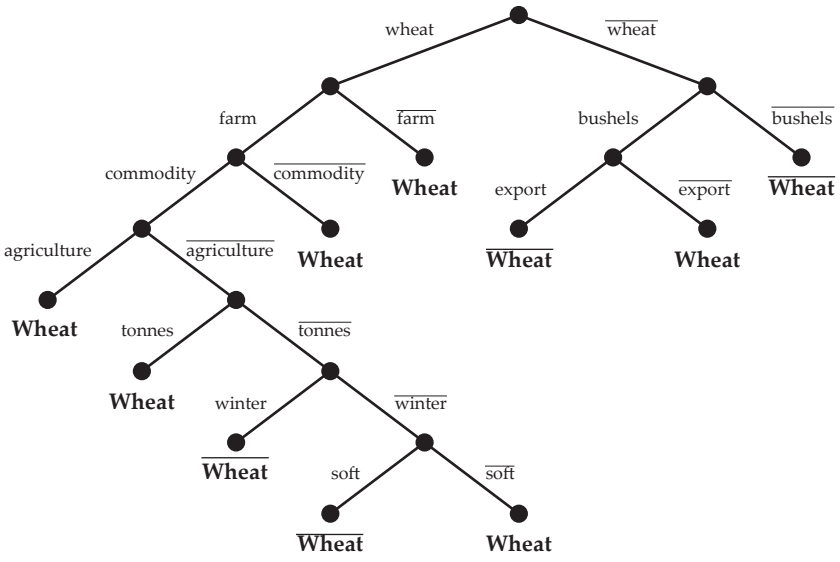


Figure 18: An example of a decision tree classifier adapted from Sebastiani [117], p. 22. Edges are labelled by features and leaves are labelled by categories (an overline indicates negation).

Decision trees are a created using a top-down approach to learning, also called the 'divide and conquer' method. This is a two stage process where (1) the learner checks to see if all the training documents

belong to the same class, and (II), if not, then it tries to find a feature that will divide the documents into two groups. It attempts to do this in such a way as to minimise the spread of classes between the two groups. The learner then repeats this process recursively for each group until if finds completely homogenous groups.

In contrast to decision trees, *inductive rule learners* use a bottom-up approach to build the set of rules. The learner starts by using every feature in the vector as a condition on every training document (e. g. Figure 19). It then goes through a process to simplify the rule set (e. g. removing redundant conditions, or merging conditions) while still maintaining the ability to classify the training set correctly. Often the learner will then go through a generalisation process to help prevent over-fitting.

IF document contains	(*programming*	AND
	Java	AND
	class inheritance	AND
	object oriented)	
AND does not contain	(*coffee*	OR
	roasting methods	OR
	arabica beans)	
THEN	classify in PROGRAMMING	

Figure 19: An example of a starting rule for a training document used by an inductive rule learner.

The advantage of the inductive rule learner is that it tends to produce a more compact representation than decision trees.

Neural Networks

Neural network classifiers are based on artificial intelligence techniques that attempt to model the human brain. Unlike other approaches, the neural network is trained through a trial and error process. The learner is given training document and 'guesses' which category it belongs to, and if the guess is incorrect, then the learner adjusts various parameters in the network to minimise error.

A simple neural network can be drawn something like Figure 20. Each of the nodes takes the feature values as inputs, and outputs a 'vote' on how the document should be categorised. Neural networks can have many more nodes and become quite complex. Because of this, they can also be expensive to run in terms of processor time. The structure of the neural network lends itself to parallel process-

ing however, which can significantly reduce processing time in some circumstances.

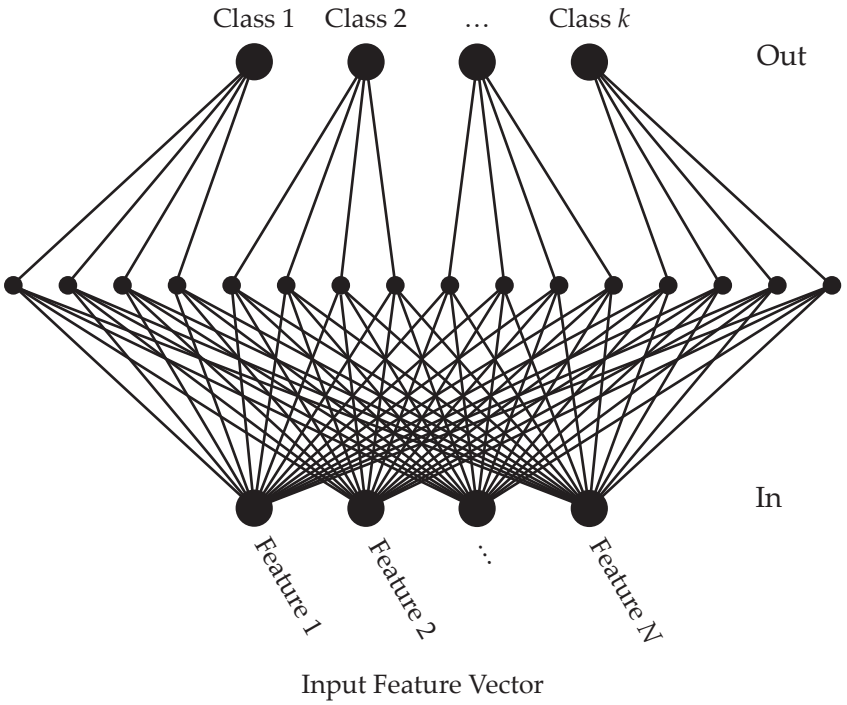


Figure 20: Example representation of a neural network, adapted from Lubbes [77], p. 64

Support Vector Machines

The *support vector machine* (SVM) method has gained increasing popularity and is relatively new compared to other classification methods. It is a mathematical technique that is perhaps best explained through an example (see Figure 21). In the diagram, crosses represent training documents belonging to the class COFFEE, and circles represent other classes. The SVM method attempts to find the widest possible parallel lines that will separate the two groups (thin black lines in the diagram). The 'best' set of parallel lines will be defined by only three training documents, called *support vectors* (marked by squares in the diagram). It then creates a dividing line (called a *decision surface*) to mark the boundary between COFFEE and everything else (marked by the thick line in the diagram).

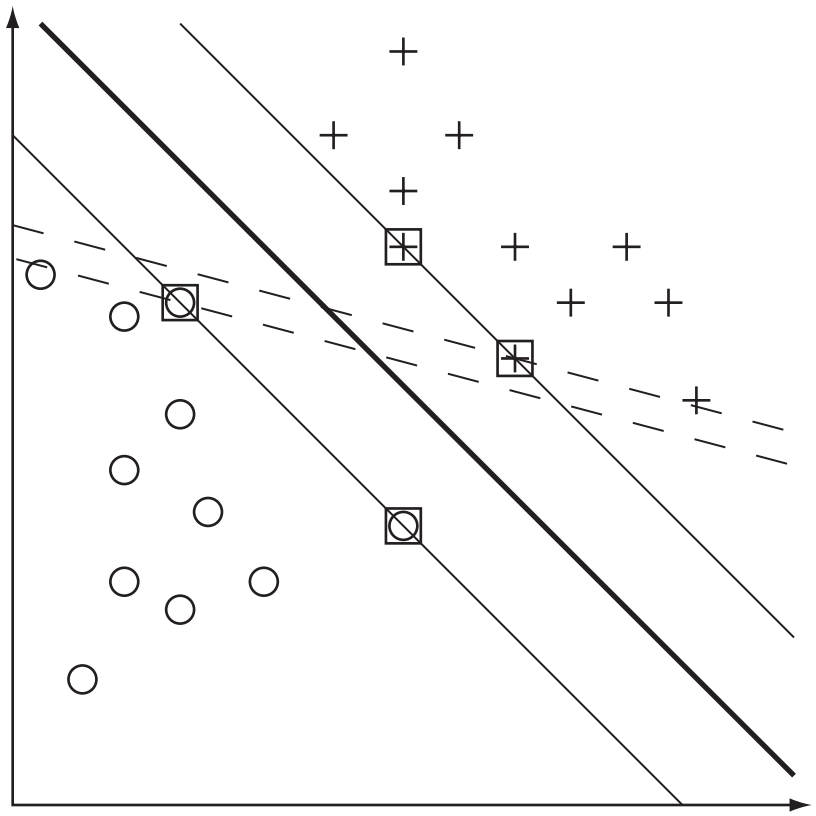


Figure 21: Example to illustrate support vector machines (SVMs), adapted from Sebastiani [117]. Support vectors are marked by squares, thin lines (dashed and solid) indicate possible divisions between the two categories, and the thick line represents the decision surface.

In our example it was possible to draw a straight line which divided the class COFFEE from other documents. In practice however, this is not always possible. In these cases the SVM method uses mathematical transformations to map the document vectors onto a dimensional space where drawing a straight line *is* possible.

6.4.3 *Advantages and Disadvantages of Automatic Classification*

The advantage of automatic classification schemes is that they can classify very large numbers of documents much more quickly than human classifiers. In some cases the cost of setting up an automatic classifier will be considerably less than employing an army of professionals to classify every item in the corpus.

Automatic categorisation can be extremely useful also; however, it does have some drawbacks. Lubbes [76] and Meyers [95] identify a number of key disadvantages with automatic categorisation:

- *Limited accuracy:* Automatic classifiers are not perfect, and will always require human intervention to find misclassified documents. In addition, 'the imprecision of an automatic system can increase the amount of time required to find information, or worse, make information altogether inaccessible, reducing productivity and restricting access to potentially critical information' [95, p. 20]. The performance of automatic classifiers is difficult to assess. Sebastiani [117] compared reported results on standard document sets for a range of different classifiers. Accuracy values ranged from 44% to 92%, with a mean of around 78%. If we make an optimistic assumption that a classifier will perform 90% of classifications correctly, this means that a human being will still need to intervene and reclassify 10 in every hundred documents. This may be a significant labour saving over having to classify all 100 documents, but human intervention cannot be eliminated.

- *Does not work well on very short documents, large documents, or on documents without uniform contents:* The classifier's effectiveness depends on how many categories are required, the size of the collection, and the nature of the documents classified. Email documents in particular, tend to be short and often do not contain enough information to be classified properly. For example, an email responding to a question may contain the single word 'yes', which on its own has very little information content. The

automatic classifier has no access to the context of the email and will place the message in a miscellaneous category.

Large documents (e. g. books) also cause problems because they tend to cover a range of different topics, making them difficult to classify. Some authors have recommended breaking such documents up into sections or chapters, for better classification, however this may not be practical in all circumstances.

Since automatic classification systems generally operate by characterising *text*, they are often not well suited to documents with large amounts of multimedia content such as photographs, diagrams or sound data.

- *Inflexibility:* Whenever the classification scheme is changed, the classifier must be re-trained, or its rule set re-defined. If a pattern matching system is used, an new set of training documents must be selected to reflect the new classification scheme. If a rule based system is used, then the rules must be redefined, requiring the presence of expert personnel.

In summary, 'humans have an advantage [over automatic classifiers] because they are able to use context—information *not contained in the document itself*—to make decisions on how the document should be classified' [95, p. 21]. In the context of the thesis aims, automatic classification systems do not really eliminate the need for a skilled administrator—someone must still be available to reclassify incorrectly classified items. The automatic classification system also requires reconfiguration whenever categories are added or removed. Automatic classification techniques may prove useful tools, but are not a complete solution in and of themselves.

6.5 UNCONTROLLED VOCABULARIES

Another way to solve the issue of requiring a dedicated administrator is the use of uncontrolled vocabularies. Librarians and information scientist have debated the merits of uncontrolled vocabularies for many years. Unfortunately, the issue is muddied because the phrase 'uncontrolled vocabulary' can mean two things: I) free text search, and II) author-supplied metadata. This results in two debates regarding 'uncontrolled vocabulary': I) Are professionally created thesauri necessary when free-text search is available? and II) Can the quality of author-supplied metadata be trusted? In this section we will examine

each of these debates in turn.

6.5.1 *Free Text Search*

If categorising information is such a difficult task, it is appropriate to ask the question 'why bother at all?' Computers are extremely good at searching through large amounts of text, and the IR community has spent enormous amounts of effort on perfecting search processes. With increasingly effective search engines, is there any real benefit to be gained from a human-generated categorisation scheme?

In the field of LIS the debate over controlled vocabulary versus free text search has been carried out for decades. Svenonius [133] traced the debate in literature all the way back to 1854, where the controversy was over whether book titles were sufficient for a card catalogue, or if a subject categorisation scheme was also necessary. Svenonius then reports that the development of successful classification schemes such as the LCSH appeared to close the debate until the development of computer technology in the late 1950's. The computer offered the promise of a 'quick and effective form of derived indexing' [133, p. 333]. Since then, the debate has continued over whether human-generated classifications are really necessary.

The main argument from the LIS professional's point of view is that a computer can never replicate the sophisticated knowledge a human brings to the problem of categorisation and indexing. In a series of two papers addressing this question, Anderson and Pérez-Carballo [7, 8] concede that modern search algorithms are indeed sophisticated and effective. They counter argue, however:

> Nevertheless, effective human indexing relies on a very sophisticated use of human intelligence. Machines are very far from simulating the work of a human indexer. Part of what a human indexer does is to interpret the text (understand the message). Human indexers do this in the context of their cultures and their personal experiences, including their prejudices, as well as taking into consideration user needs and desires. [7, p. 238]

A number of studies have been performed attempting to compare the effectiveness of controlled vocabulary versus free text search. Fidel [38], for example, conducted a study of an online library catalogue and found that searchers tended to use both controlled vocabulary terms and free text searches to find documents. 'The nature of a term

and of a request, as well as the searcher's personal preference, contributed to the selection of search terms, but the number of databases required for a request, the availability and quality of the thesaurus, and the quality of indexing were the major factors to affect search-term selection' [38, p. 1]. Since then, a number of studies and reviews [e. g. 47, 114] have found similar results: that people tend to make use of both controlled and uncontrolled vocabulary in locating documents.

With the growth of the internet and the success of search engines like Google however, people are becoming more and more familiar with search engines. Barnum et al. [9] conducted a study comparing free-text searching with a hyper-linked index, asking participants to perform information-seeking tasks using an electronic (PDF) version of a book. The study found that people preferred the full-text search tools, even though they performed tasks quicker, and more accurately using the index. These results are similar to findings in Chapter 7 of this thesis, where people stated a preference for full-text search simply because they were 'used to it'.

The disadvantage of free-text search is that while it tends give higher recall, it provides low precision [141]. That is, the free-text search will return many more documents than a controlled vocabulary search, but the results will usually be less relevant. This means that the searcher must do more work in determining which results are relevant and worth investigating.

Regardless of whether human-created indexes and categorisation schemes are 'better', there is a definite consumer demand for search engines. Anderson and Pérez-Carballo [7], p. 238 make this point clearly:

> If research into the merits of automatic vs human indexing has been inconclusive, the actual experience of IR database producers is persuasive. The fact that IR databases that rely solely on automatic indexing have been economically successful means that the users who are paying for them (either in actual financial outlay or in time spend using them or both) find them sufficiently effective to justify the cost. In some situations, no other options are available.

It seems that search engines are here to stay.

A great irony in this debate is that Google was successful precisely because it attempted to use human-generated metadata in addition to

text analysis. The PageRank [17] algorithm examines hyperlink structure on the web, making the simple assumption that 'a hyperlink from page A to page B is a recommendation of page B by the author of page A' [54]. In making this assumption, the PageRank algorithm is reliant on human beings making intelligent decisions in linking one web page to another. In doing so it is making implicit use of the *context* for a web page. Hence, humans can never be taken out of the equation.

Is free text search then, the solution to the categorisation problem? Free-text search engines have certainly become very effective tools for information retrieval. They are limited however, since they are reliant on *text*. One of our requirements listed in Chapter 5 was that the classification system be suitable for multimedia data. Like automatic classification, free-text search does not generally cope well with very short or very long documents; hence, while free-text search is a useful tool, on its own it does not constitute a complete solution.

6.5.2 *Author-Supplied Metadata*

The other side to the uncontrolled vocabulary debate is author-supplied metadata, that is, metadata (including categorisations) supplied by content authors rather than information professionals. Often this is supplied in the form of subject keywords. This kind of metadata is sometimes used by scientific journals to supplement indexing, and the use of the 'keywords' meta-tag was an early feature of the WWW [98]. By allowing content authors to categorise their own work, categorisation becomes a collaborative task, rather than the burden of a few individuals.

Using keywords in this way is less costly than employing an expert to perform a domain analysis to determine an appropriate classification scheme. Allowing authors to classify their own work also avoids the ongoing cost of employing an expert to classify resources added to the repository. There are problems however, with allowing authors control over classification.

Firstly, there is the question of expertise, an issue that is summarised well by Barton et al.:

> The basic problem is this: who is best placed to add subject-related metadata for maximum resource discovery, the author who may know the subject area and its terminology well, or a metadata specialist, who may be better placed

> to step back and think about all the potential users of a resource, or about consistency of subject terms and classifications across a repository, archive or network, so that like really is classified with like. [12, p. 3]

Some researchers like Greenberg et al. [46] argue that authors are quite competent at creating metadata, and 'in some cases they may be able to create metadata that is of better quality than a metadata professional can produce'. However, even if authors are competent at creating metadata, there is the question of why they bother to do so.

Secondly, there is the issue of deliberate misuse of metadata. There is a perception that author-generated metadata lacks credibility. Thomas and Griffin [11] argue that there is little incentive for authors to put time and effort into providing quality metadata unless it is for the purposes of self-promotion. Thus, author generated metadata is perceived as inaccurate at best, and deliberately misleading at worst [33, 135].

In an enterprise context, the problem of self-promotion may be lessened, however the questions of quality and motivation still remain. In many cases, there is very little incentive for authors to provide quality metadata. Often the people contributing the metadata are not the same people who use the information (as seen in Chapters 2 and 5). Hence, amateur categorisation does not seem to be a viable solution to the categorisation problem unless mechanisms can be found to improve the quality of the data and give people an incentive to provide it.

6.6 FOLKSONOMIES

The term *folksonomy* describes the dataset that results when people use a tagging service. A tagging service is one that allows participants to associate keywords (called *tags*) with resources (usually via the internet or intranet). The resource may be almost anything. Tagging services exist to tag an enormous variety of things such as photographs, URLs, podcasts, computer games, music and videos. In recent years, tagging services such as Delicious[†] (for tagging URL bookmarks) and Flickr[‡] (for tagging photographs) have become increasingly popular. For an overview of tagging services, we refer the reader to an excellent review by Hammond et al. [50].

† http://del.icio.us/
‡ http://www.flickr.com

A folksonomy places very few (if any) restrictions on which keywords are permissible. Authors often contrast this to other methods of categorisation that are hierarchical and have mutually exclusive categories [120, 143]. Using an uncontrolled vocabulary in this way leads to a number of well-documented problems (see Section 6.5). A folksonomy-based system sidesteps the issue of author vs professional by placing responsibility for categorisation in the hands of people using the system. In fact, the term folksonomy is a portmanteau of folk (meaning people) and taxonomy, in other words, 'taxonomy of the people'.

Another distinct feature of tagging services is that tagging is done in a social context. In general, a tagging service will make the tags and resources contributed by one person visible to everyone else using the system. For example, in Delicious (a popular service for tagging URLs), I might tag the URL http://www.anu.edu.au with 'ANU', 'university' and 'research'. These tags are made visible to others, and I am able to view tags others have used, such as 'education' and 'Australia'. I am also able to see all the other websites people have marked with 'ANU' or 'research'. The sharing of tags allows people to discover others with similar interests, and to change their tags in response to others' behaviour. This is a key aspect of tagging services that separates them from other methods of information organisation.

6.6.1 *Advantages of Folksonomies*

The distinctive properties of folksonomies result in a number of advantages over other methods of information organisation. Mathes [86] and Kroski [66] outline many of these, which I have summarised under six headings:

Serendipitous browsing

Folksonomies provide excellent support for browsing strategies and discovering resources unexpectedly in a general area of interest. Marchionini [83] writes:

> 'In general, browsing is an approach to information seeking that is informal and opportunistic and depends heavily on the information environment. [...] Browsing is particularly effective for information problems that are ill defined or interdisciplinary and when the goal of information seeking is to gather overview information about a

topic or to keep abreast of developments in a field.'

Given that most real-world problems tend to be both ill-defined and interdisciplinary, browsing is an important method of information seeking.

Tagging services support browsing by providing the ability to view data from multiple perspectives. Millen et al. [96] call this *pivot browsing* and describe it as follows:

> 'A number of user interface elements allow social browsing of the [resource] space. For example, user names are "clickable" links; clicking on a name reveals the [resource] collection for that user. This allows someone to get a sense of the topics of interest for a particular user. Similarly, tags are also clickable, and when selected will result in a list of all [resources] that share that tag. This is a useful way to browse through the entire [resource] collection to see if it includes information sources of interest.'

By viewing the resource collection of another person who shares similar interests, I may discover important resources I would not have otherwise come across. In this way, tagging services provide many opportunities for discovering resources in addition to full-text search methods.

Accurate reflection of user vocabulary

A folksonomy 'directly reflects the vocabulary of the users' [86]; there is no need for tag creators to translate between their own vocabulary and that of the system designer. Merholz [91] likens folksonomies to 'desire lines', the 'foot-worn paths that appear in a landscape over time. [. . .] [T]hese trails demonstrate how a landscape's users choose to move, which is often not on the paved paths'. As users add tags to a folksonomy, the popular tags emerge as a consensus view on how a resource should be tagged.

The social aspects of folksonomies reinforce this, as people tend to modify their vocabulary to match the usage of others around them. The term folksonomy itself is a good example of this. Merholz's article on desire lines does not contain the word folksonomy, and he has written on his personal website that he thinks the term is incorrect and misleading [92]. On Delicious however, people have tagged the article with folksonomy 31 times (as of 29 January 2007), as folksonomy has become accepted as a consensus label.

Low barriers to entry

The lack of restrictions on keywords means that people don't need to learn a complicated classification system to use a tagging service. Users require no special training or previous knowledge to participate. Typing free keywords is simpler and easier than determining the correct location for a resource in a complex hierarchy. Because of this, people are much more likely to contribute this kind of metadata, as opposed to more involved schemes such as the Dublin Core. Tagging is simple and easy.

Social feedback

The ability to see how other users tag similar items provides a feedback mechanism that encourages consensus while allowing for disagreement. For example, I might tag a website such as http://www.anu.edu.au with 'universities'. After I enter my tag however, I discover that most other people use the singular 'university', so I change my tag to match theirs. On the other hand, people may use a differing tag when they mean something specific-such as 'cinema' versus 'film' [120]. The social feedback in a tagging service encourages consensus, but does not impose it.

Intrinsic incentives

Folksonomies work because the ability to tag items is personally useful for organising information. For example, the Delicious website allows people to store bookmarks on a central server. This means that a user's bookmarks are available from any PC connected to the internet. Tags allow bookmarks to be organised so that they can easily be found again later. Even if a person never browses anyone else's bookmarks, the system is still useful. Thus tagging has a personal incentive rather than being an optional extra solely for the good of the community.

That said, folksonomies do provide a social encouragement to tag. If a person finds a particular resource valuable, then tagging allows them to share it with others who have similar interests. There is also the knowledge that tagging a resource benefits the rest of the community. By tagging a resource and adding to the collection, I am making it available to others who may also find it useful.

Flexibility

The unstructured nature of folksonomies allows for unanticipated uses. For example, one of the more popular tags on Delicious is 'toread' (for 'to read'). Obviously, this is of little use for describing what a website is about, but individuals find it useful to mark items for later perusal. The presence of the 'toread' tag does not disrupt the operation of the system. Another example is the use of *geotagging* in Flickr, a popular photo tagging service. Geotagging is where someone tags a photograph with the geographical coordinates of where it was taken. The unstructured nature of tags allows great flexibility in how they are used.

The lack of restrictions on tags also means that folksonomies have the ability to adapt to changes in vocabulary and usage. It costs very little (if anything) to add a new term to the vocabulary, and items can be co-tagged as the vocabulary changes.

6.6.2 *Disadvantages of Folksonomies*

Uncontrolled vocabularies have a number of well-known problems [133] including synonymy, homonymy and noise.

Synonymy occurs when a number of words (or tags) share the same meaning. For example, to describe a photograph of a Macintosh Computer, I might use the tag and 'Macintosh'. Another person might use the tag 'mac', while another uses the tag 'apple', all referring to the same thing. If I browse the folksonomy using only one of these tags, I might miss important resources that have been tagged using only a synonym. In information retrieval terms, synonymy leads to poor recall.

Another perennial problem with uncontrolled vocabularies is that of homonymy-where a single tag has multiple meanings. 'Apple', for instance, may mean an Apple computer, or a type of fruit. A search for 'apple' to find resources related to fruit may return a number of irrelevant resources related to apple computers. In this regard, acronyms such as 'AJAX' or 'ANT' can be particularly troublesome. In information retrieval terms, homonymy leads to poor precision.

Tag noise reduces precision in folksonomies even further. There are no checks on the tags people enter; some tags may be misspelled, while others may be just plain wrong. As mentioned above, tags such as 'toread' tell us very little about a particular resource. Some people may try to improve the popularity of a particular resource by assign-

ing it many irrelevant, but popular tags. Compared to a professionally controlled hierarchical classification scheme, folksonomies are messy.

To summarise, while folksonomies have a number of distinct advantages arising from their collaborative, unstructured nature, they are not without their drawbacks. Not every information management problem is suited to a folksonomy-based solution.

6.7 CONCLUSION

In terms of the thesis problem, none of the approaches described above completely addresses all of the requirements. Each of the approaches described have a number of drawbacks and will not work well in all situations (see Table 27). In spite of this, the *folksonomy* approach appears to have the most potential to address the thesis requirements. The following chapters of this thesis address how folksonomies may be adapted to suit the specific problem we are trying to address.

Table 27: Summary of advantages and disadvantages for approaches in the literature

ADVANTAGES	DISADVANTAGES
Ecological Approaches	
Accurately reflects user vocabulary and well suited to the specific situation.	More methodology than method. Does not provide mechanisms to deal with category change. Time consuming to carry out.
Faceted Analysis	
Allows corpus to be viewed from multiple perspectives. Changing one facet does not affect other facets.	Requires expert administrator to develop and maintain the classification scheme. Does not really provide mechanisms to deal with change.
Card Sorting	
Accurately reflect users' mental models. Simple, cheap and effective.	Doesn't really specify how to structure categories. Does not provide mechanisms to deal with change. Requires expert administrator to develop and maintain the category scheme.
Automatic Categorisation	
Saves on labour costs. Can categorise many documents very quickly.	Limited accuracy. Requires expert intervention if categories are added or changed. Not well suited to multimedia data.
Uncontrolled Vocabularies—Free Text search	
Does not require an expert administrator. People are familiar with searching this way.	Not well suited to multimedia data. Can give low precision (depending on search engine).
Uncontrolled Vocabularies—Author-supplied metadata	
Does not require an expert administrator. Cheap.	Perceived as inaccurate and unreliable. Difficult to motivate authors to provide good quality data.
Folksonomies	
Reflects user's vocabulary. Responds well to change. Does not require an expert administrator. Well suited to multimedia data.	Synonymy, homonymy and tag noise. May not work well with small data sets.

Part II

FOLKSONOMIES

In Chapter 6 we examined potential solutions to the problem posed by this thesis, concluding that *folksonomies* appeared to be the best approach. Folksonomies have a number of advantages over other approaches:

1. *Folksonomies directly reflect user vocabulary*. Categorisation in a folksonomy is performed by people who use the system, as opposed to authors or librarians, so the vocabulary in a folksonomy will always reflect the language of people using the system. The vocabulary is not imposed by an epistemic authority, but emerges from the consensus views of taggers.

2. *Folksonomies are able to adapt to changes in vocabulary*. As peoples' vocabulary changes, the vocabulary in the folksonomy will change with it, since it is the users of the system who are tagging. There is no need for expert human intervention.

3. *Folksonomies do not require a dedicated administrator*. Because tagging is performed by people using the system, no dedicated librarian is required to perform classification work. Furthermore, because the category structure created by tags is dynamic, administrators are not needed to continually maintain the system.

4. *Folksonomies are suitable for multimedia data*. The success of tagging websites such as Flickr[§] and Last.fm[¶] clearly demonstrate that folksonomies are well suited to multimedia data.

Because of these advantages, folksonomies appear to have the most potential to fulfil the requirements set out in Chapter 5. Folksonomies are not without their drawbacks however, as explained here:

- Tagging services and folksonomies are a new approach, with papers in the literature appearing only since 2004 [e. g. 50, 86]. Because of this, there are many aspects of folksonomies that have not been thoroughly researched and tested. There are many open research questions such as: 'How can folksonomy data best be used to facilitate access to resources?' and 'What motivates people to choose the tags that they do?'

§ http://www.flickr.com/
¶ http://last.fm/

- Folksonomies also exhibit the well documented problems of an uncontrolled vocabulary (as discussed in Chapter 6). How to address these problems in folksonomies is still an open area of research.

- Because folksonomies operate based on emergent consensus, they work best when lots of people contribute tags. This means that they may not be well suited to smaller-scale systems. To our knowledge however, no research appears to have been carried out to determine how many people are needed to make a folksonomy work well.

The next two chapters report on studies investigating these issues further. Chapter 7 reports on a study investigating user interfaces for folksonomies, examining the 'tag cloud' and how it may be used to facilitate retrieval. Chapter 8 reports on a study comparing different clustering algorithms for folksonomies. Clustering tags can be used to address the problems of homonymy and synonymy in tag clouds, and may also make them more suitable for smaller scale information systems. Folksonomies are an exciting area of research with many important questions still unanswered.

7

INVESTIGATION OF FOLKSONOMY TAG CLOUDS

As discussed previously, there are many unanswered questions related to the use of folksonomies in information systems. One in particular is how best to present folksonomy data to promote resource discovery. To investigate this issue, I conducted an investigation into the use of weighted lists as a tool for accessing tagged information.

The weighted list, known popularly as a 'tag cloud', has appeared on many popular folksonomy-based websites. Flickr, Delicious, Technorati* and many others have all featured a tag cloud at some point in their history. However, it is unclear whether the tag cloud is actually useful as an aid to finding information. In this experiment, the system gave participants the option of using a tag cloud or a traditional search interface to carry out information seeking tasks. The results indicated that where the information-seeking task required specific information, participants preferred the search interface. Conversely, where the information-seeking task was more general, participants preferred the tag cloud. While the tag cloud is not without value, it is not sufficient as the sole means of navigation for a folksonomy-based dataset.

7.1 INTRODUCTION

A characteristic feature of folksonomies is that they generally make a participant's tags and resources visible to other participants. This makes it possible to browse friends' and colleagues' collections of resources. For instance, if I know that Bob has done a lot of work with XML, then I may browse his resources tagged *XML* to look for a good introduction to the topic. Because I know Bob is likely to have a good knowledge of this area, I assume that the resources he has tagged will be of good quality.

Given the social aspects of folksonomies, there is clearly potential for these services to support knowledge sharing and collaborative work within the enterprise. The sharing of key resources and the generation of emergent vocabulary may well help decrease the cognitive distance between people working in an organisation. This helps to

* http://technorati.com/

promote social capital, a key factor in knowledge sharing [61].

Although there is clear potential for folksonomies in the enterprise, there are still key questions still to address. Some of these include:

- What factors are important to consider when scaling down a tagging system from the whole-of-internet, to small- or medium-sized organisation scale
- How can we present folksonomy data in ways that enhance peoples' ability to locate relevant information?

In this chapter, we examine aspects of these questions with particular respect to user interfaces for folksonomies.

7.1.1 *Tag Clouds*

Tag clouds are a user interface element commonly associated with folksonomy datasets (see Figure 22 for an example). The Wikipedia encyclopaedia contains the following entry regarding the tag cloud:

> 'A tag cloud (more traditionally known as a weighted list in the field of visual design) is a visual depiction of content tags used on a website. Often, more frequently used tags are depicted in a larger font or otherwise emphasized, while the displayed order is generally alphabetical. Thus both finding a tag by alphabet and by popularity is possible. Selecting a single tag within a tag cloud will generally lead to a collection of items that are associated with that tag.'†

In essence, the tag cloud translates the emergent vocabulary of a folksonomy into a social navigation tool [32].

Although tag clouds are extremely common among tagging services, the academic literature appears to be largely silent on the subject. In the blogosphere however, there has been heated discussion amongst website design and information architecture practitioners. Jeffrey Zeldman [148] initially criticised tag clouds simply for being popular. However, in the same article he described tag clouds as being 'smart' and said that this smartness is the reason 'why so many have rushed to use them'. In a subsequent article he criticised tag clouds as

† 'Tag Cloud', *Wikipedia Encyclopedia*. Retrieved 17 July 2006.
URL: http://en.wikipedia.org/w/index.php?title=Tag_cloud&oldid=63880649.

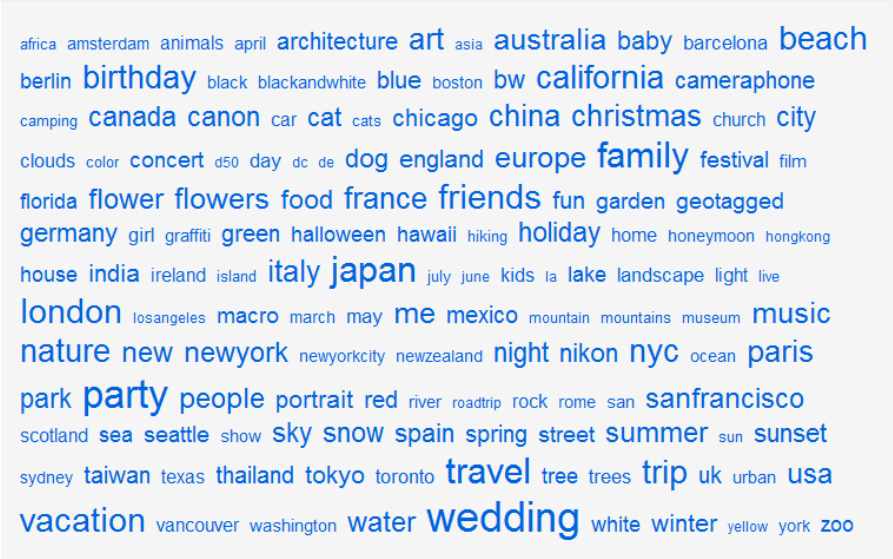


Figure 22: An example of a tag cloud (taken from http://www.flickr.com/photos/tags/ on 16 June 2007).

obscuring useful information by being skewed towards what is popular. He claims that tag clouds 'create a false intellectual equivalence between high- and low-level topics' [147]. Some of his criticisms were not so much of tag clouds, however, as folksonomies in general.

Some bloggers, such as Jon [65], wrote responses to these articles and defended the tag cloud as a navigation aid. Jon claimed that the tag cloud is a powerful tool that has yet to realise its full potential. Others, such as Porter [106], responded to the broader criticisms of folksonomies in general. Zeldman appears to claim that the popularity metric hinders access to useful items that are not tagged with popular keywords. Porter counter-argues however, that just because an item is popular, that does not necessarily mean it is irrelevant. Notably, Tomas Vander Wal—the man credited with coining the term folksonomy [139]—described tag clouds as being 'cute but offer little value' [140] . While the discussion in the blogosphere has been lively, very few of the authors have performed empirical studies to verify their claims. As Brooks and Montanez point out:

> [M]uch of the discussion regarding the advantages and disadvantages of tags and folksonomy has taken place within the blogosphere, as opposed to within peer-reviewed conferences or journals. The blogosphere has the great advantage of allowing this discussion to happen quickly and provide a voice to all interested participants, but it also presents a difficult challenge to researchers in terms

> of properly evaluating and acknowledging contributions that have not been externally vetted. [18]

This lack of empirical evidence prompted this investigation of the advantages and disadvantages of tag clouds for exploring folksonomy data. This fits in with the broader aim of investigating the use of tagging systems in smaller scale information systems.

7.1.2 *The Study*

The question addressed here was 'Do tag clouds provide value to people seeking information from a folksonomy data set?' To answer this question I created a folksonomy-like dataset, and asked participants to perform information-seeking tasks on it. The system provided participants with the option of using either a tag cloud or a search box interface to query the database. The hypothesis was that if people find no value in a tag cloud for information seeking, then they would primarily use the search box to seek information. At best, people would attempt one or two queries using the tag cloud, then resort to the search box.

The results indicated that while some participants preferred to use the search box exclusively, a significant proportion of participants used the tag cloud to find information. In total, the number of questions answered using the tag cloud was greater than those answered using the search box. Thus, we conclude that the tag cloud does provide value for navigating a folksonomy dataset.

Examining survey comments and usage patterns of the data, we explore the benefits and limitations of the tag cloud.

7.2 METHOD

A key characteristic of the folksonomy is that it is user created metadata [86], that is, the people who use the system are also the ones tagging items. Hence, instead of downloading an existing folksonomy data set for use in this study [as with other studies such as 25, 45, 73], I asked participants in the study to tag articles from Slashdot Science[‡] to create a folksonomy-like data set. This ensured that the participants querying the dataset had also contributed tags to the dataset.

‡ http://science.slashdot.org/

To conduct the experiment, I recruited participants from students enrolled in a first-year engineering class offered by the Department of Engineering at the ANU. Table 28 summarises other characteristics of the participants.

Table 28: Participants

Participants:	89
Female:	6
Male:	83
Age range:	18–24
Participants with a blog:	10
Participants who use a social bookmarking service (e. g. Delicious, Furl, Magnolia) :	2
Articles tagged:	973
Mean articles tagged per session:	108.1
Mean articles tagged per participant:	10.9

In the experiment, I asked participants to tag around ten articles each, using the interface shown in Figure 23. The system selected articles randomly from a pool of 1074 articles, dated from 17 August 2005, to 25 May 2006. Explaining the experiment and allowing participants to tag ten articles took approximately 25 minutes for each session.

The system allowed participants to enter as many tags as they wished, separated by commas. This made it possible to use spaces in tags, rather than restricting the participant to a single word. This is in contrast to many popular tagging systems such as Delicious and Flickr, which only allow single word tags. Allowing the use of multi-word tags eliminated the problem of establishing a convention for word combination. One study [85] found that at least 10% of tags in Delicious were compound words created 'by putting a symbol or piece of punctuation inside the tag to represent a space', yet no consensus or convention had been chosen by the Delicious user community. Allowing spaces (in a manner similar to that adopted by RawSugar[§]) avoids this issue.

The tagging interface (Figure 23) also listed all the tags a participant had used to date on the right-hand side of the screen under the heading 'My tags'. Clicking on one of the tags added the tag to

§ http://www.rawsugar.com

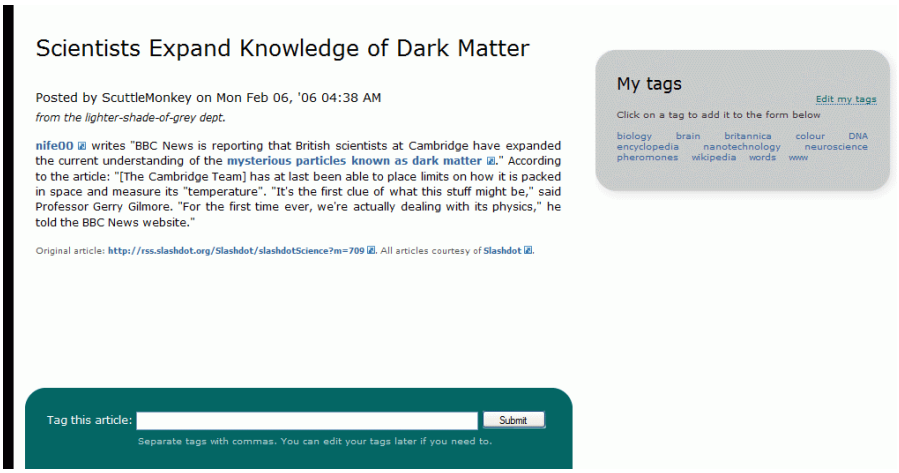


Figure 23: The interface for tagging articles.

the form at the bottom of the screen. This is a timesaving feature common in tagging services such as Delicious and CiteULike¶. I included this feature to encourage consistency in tagging by reminding participants of tags they had used previously. For example, a participant might enter 'web design' for one article and 'website design' for another. Including the 'My tags' panel helped to limit this individual non-conformity.

Each session was divided into two halves, with the first half being devoted to tagging articles, and the second half devoted to querying the database of tagged articles. In the second half, I asked participants to answer ten questions (see Table 29) by querying the collection of tagged articles. The questions were designed to elicit a range of information-seeking behaviours, ranging from general browsing to specific information seeking. Questions 2 and 4 were the most general, giving no specific topic to search on, while questions 1 and 6 related to specific topics represented by a relatively small number of articles in the data set.

To query the database, the system presented participants with the option of using either a tag cloud, or a conventional search box. Both options were presented on the same screen, with approximately half of the horizontal screen space being given to each (see Figure 24). The system recorded which interface participants used to make their queries, and which question was displayed on the screen at the time.

The tag cloud displayed the top 70 most-used tags in the database. As folksonomy frequency tends to follow a power law distribution [25], I made the size of each tag proportional to the log of its fre-

¶ http://www.citeulike.org

Table 29: Questions asked of participants to elicit information-seeking behaviour.

1.	Tell me about an interesting application of nanotechnology.
2.	Paste the title of an article you find interesting.
3.	Paste the title of an article that contains something about Mars.
4.	Paste the title of an article you find amusing.
5.	Paste the title of an article related to biology.
6.	Tell me something about sustainable fuels.
7.	Tell me about a scientific idea/discovery that has military applications.
8.	Paste the title of an article related to NASA.
9.	Paste the title of an article related to archaeology.
10.	Paste the title of an article related to video game addiction.

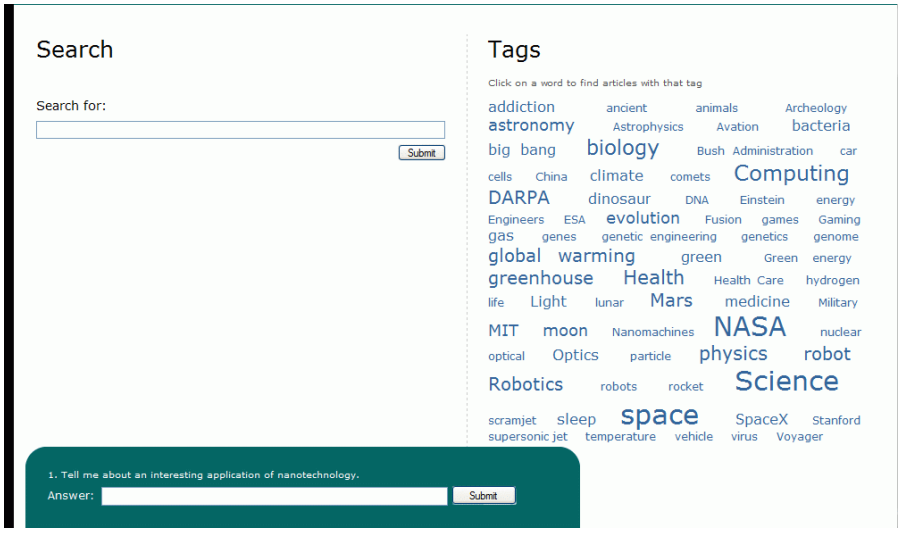


Figure 24: The search interface showing the search box and tag cloud. This particular screenshot is from the third session.

quency. The specific equation used was:

$$\text{TagSize} = 1 + C\frac{f_i - f_{min} + 1}{f_{max} - f_{min} - 1} \tag{7.1}$$

Where f_i is the frequency of tag i, f_{min} is the minimum frequency in the top 70 tags, f_{max} is the maximum frequency in the top 70 tags, and C is a scaling constant that determines the maximum text size. This gives a tag size relative to the default font size set by the browser.

Clicking on a tag returned a list of all articles tagged with that keyword. The results page for the tag cloud interface showed a subsequent tag cloud for all the articles featuring the selected tag (Figure 25). Clicking on a tag in this sub-cloud would further refine the results to articles containing both tags, and so on for sub-sub-clouds, etc.

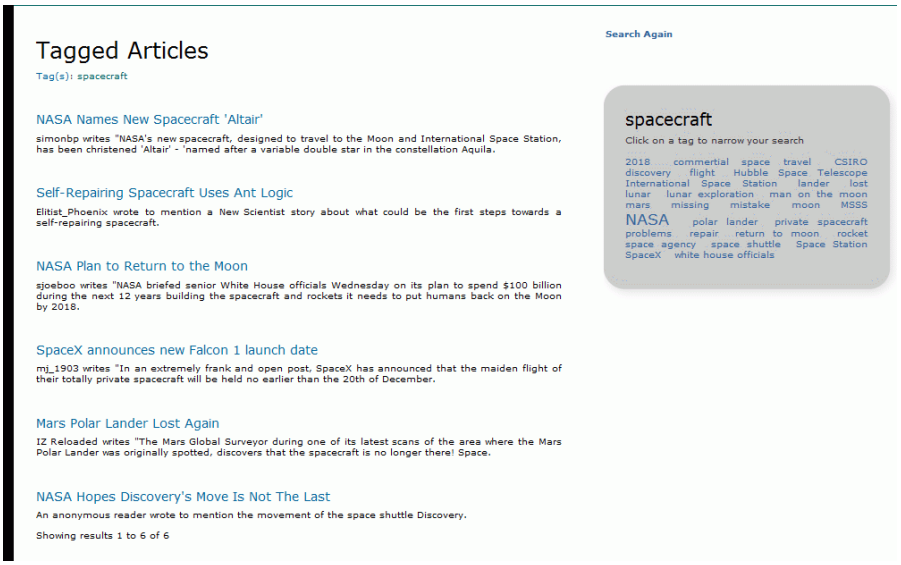


Figure 25: Example of results from a tag cloud query.

The system generated query results for the search box using the Ranking with Vector Spaces algorithm provided by MySQL‖, using the full text of each article. This algorithm uses a variation on the well known TF-IDF term-weighting method. In this method the database system assigns each term in the collection a weight, according to its frequency in the collection and its frequency in each article. When a participant performs a query, the algorithm selects and ranks articles according to the matched query terms and their individual weights.

By contrast, query results from the tag cloud were generated by

‖ See http://dev.mysql.com/doc/internals/en/full-text-search.html for details.

simply retrieving all articles tagged with the chosen keyword. That is, any article tagged with the relevant keyword was included in the search results. This however, left the question of how to rank the search results. In order to maintain consistency, the system ranked the result set displayed by the tag cloud using the same Ranking with Vector Spaces algorithm used for the search box. So, while the methods used to *select* which articles to display were different for the search box and tag cloud, the same method was used to *rank* the displayed results.

While most folksonomy-based websites order results by how recently items were tagged and the number of people tagging the item, in this system such an organisation would have had little relevance. Other methods of ranking folksonomy search results have been proposed [e. g. 60]; however, my intention was not to compare the precision and recall of the two techniques, but rather to examine user perceptions and usage patters. Hence, I attempted to maintain consistency by ranking results using the same algorithm for both query methods.

After completing ten questions, the system presented participants with an exit survey asking two questions (see Figure 26).

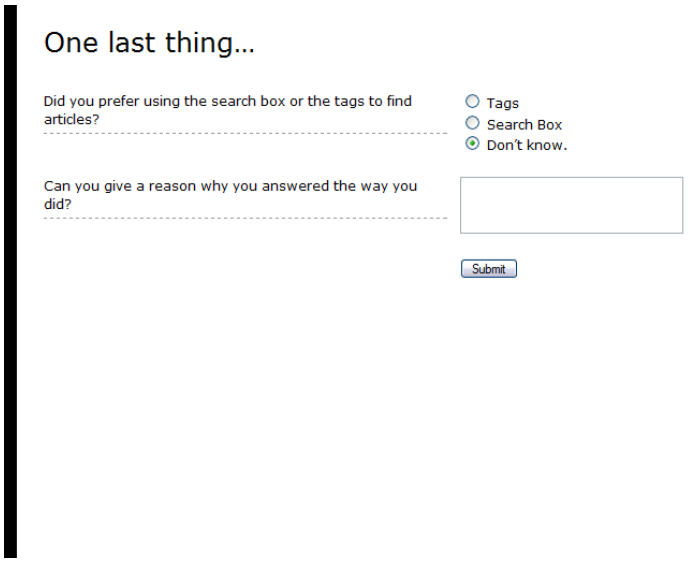


Figure 26: The exit Survey.

7.2.1 *Characteristics of the dataset*

The folksonomy dataset created through this process is quite different to that of frequently studied, large-scale systems such as Flickr

or Delicious. The experimental conditions here, however, reflect situations that often occur in an enterprise context. Hence, the study contributes to the overall thesis aims. Not all organisations have thousands of employees. Managers often delegate data entry and information management tasks to subordinates. There is potential in many organisational settings for people to be tagging resources that they did not select, without any personal benefit to themselves. Of course, such situations would not be ideally suited to a folksonomy; in practice however, all methods of categorisation are less than ideal (as seen in previous chapters).

The specific differences, in contrast to a larger-scale folksonomy, are as follows:

- *The number of participants was relatively small*, when compared with studies of tagging systems in the enterprise [such as 29, 36, 96] which report upwards of 300 participants.
- *Participants did not select articles to tag*. This is in contrast to most tagging services where participants select for themselves which resources they wish to tag. This provides a kind of quality filter not present in this experiment.
- As the 'shared' dataset was artificially selected, *participants lacked a context for tagging*. Participants in most tagging systems find some kind of personal benefit for tagging. Delicious, for example, allows participants to better organise their own bookmarks. Thus, tags created in this context have personal relevance for the participant. In this case participants were tagging simply because it was part of the experiment.
- *There was no social feedback aspect to tagging*. Unlike most tagging systems, participants were not able to browse others' tags in this experiment. Since participants did not select which articles to tag, this would have had little relevance. However, this prevented participants from observing how others tagged articles, potentially resulting in less convergence on tag usage.

These conditions produce something of a worst-case scenario folksonomy. However, if a tag cloud shows value for navigation in a less-than-ideal dataset such as this, then we assume that it will perform even better with more typical folksonomy data.

7.3 RESULTS

7.3.1 *Last-strike queries*

Do tag clouds provide value to people seeking information from a folksonomy data set? To examine this question, we assume that the last query made before answering a question provided the relevant information to answer. We shall refer to the last query made before answering as a 'last-strike query'.

In practice, this assumption about last-strike queries may not always hold true. Someone may continue querying a database, even after they have found a relevant resource, simply to see if a 'better' resource is available. In the context of this study however, the majority of questions asked participants to 'copy and paste' from articles. This makes it more likely that the last-strike query did indeed provide the relevant answer.

Each of the 89 participants answered 10 questions, giving 890 last-strike queries. We assume that those who did not search to answer a question either answered from their own knowledge, or used the results from the previous search.

Table 30 shows that participants answered the majority of questions using the tag cloud. Thus, it seems that the tag cloud does provide some kind of value. Examining the last-strike queries for each question however (see Figure 27), participants favoured the search box for six of the ten questions. Investigating this further we find two scenarios where use of the tag cloud outweighed use of the search box. The first was where the information-seeking task was broad and non-specific, as in questions 2 and 4. The second was when the tag cloud contained a keyword relevant to the question.

Table 30: Last-Strike Queries

Method:	Tag Cloud	Search Box	Did not Search
Last-strike queries:	427 (48.0%)	367 (41.2%)	96 (10.8%)

7.3.2 *Browsing versus finding*

Questions 2 and 4 were broad, non-specific questions:

2. *Paste the title of an article you find interesting.*
4. *Paste the title of an article you find amusing.*

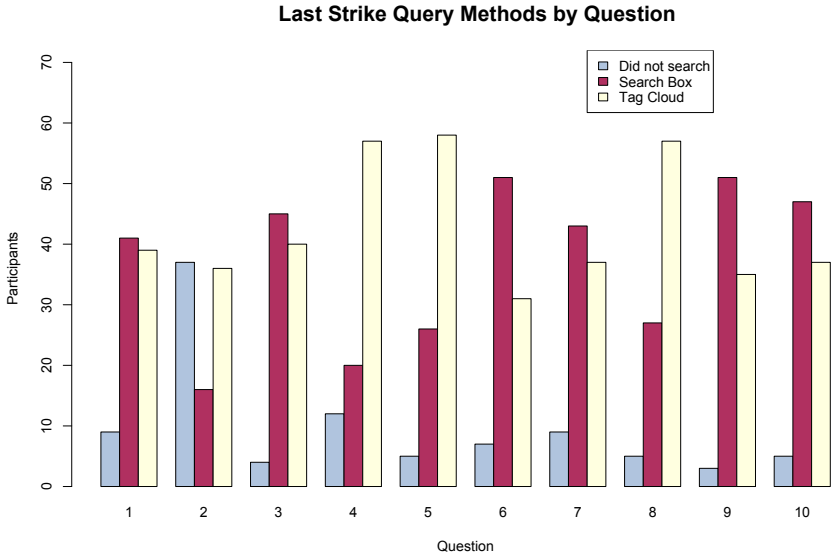


Figure 27: Query method used to answer each question

The number of laststrike queries for the tag cloud was much greater than the search box for both these questions. This suggests that tag clouds are useful when the information-seeking task is non-specific. That is, tag clouds support browsing or serendipitous discovery.

Other authors have observed that this is a feature of folksonomies in general. Mathes [86] commented that one of the strengths of folksonomies is serendipity: 'browsing the system and its interlinked related tag sets is wonderful for finding things unexpectedly in a general area.' Brooks and Montanez [18] also found that tagging is most effective at placing articles into broad categories, but is less effective as a tool for indicating an article's specific content.

In addition, participants' comments from the exit survey also support this observation. Eight respondents indicated that their choice of interface depended on the information-seeking task, often commenting that the tags supported more general browsing. Ten participants commented that the search box allowed for greater specificity when entering queries. Examples included:

> *[T]he tags were more usefull [sic.] for finding things about general topics. [T]he search bar was better for spicific [sic.] things*
>
> *[T]he search box lets you find exactly what you want, however if you wanted to find related articles, the tags are more useful*
>
> *If searching for something specific you can specify more constraints about the subject so less results are returned, otherwise*

for more general use the tags would be good

Conversely, it seems that the tag cloud was less useful when the information seeking task was more specific.

In general, it appears that answering a question using the tag cloud required more queries per question than the search box. We cannot directly compare the mean number of searches for each interface however, since a participant may have used a mixture of both methods. To compare the number of searches to find an answer, we examine the cases where a participant relied solely on one interface to answer a question.

The histograms for both interfaces (Figure 28) show that most questions were answered with a single query. In general however, participants who relied solely on the tag cloud made more queries when answering a question, even when a relevant keyword was present in the tag cloud (see Table 31). The 'tail' of the tag cloud histogram in Figure 6 shows two cases where participants made thirteen to fourteen queries before answering a question. This suggests that participants who used the tag cloud were more likely to engage in browsing behaviour.

Table 31: Mean queries to answer question where participants relied on a single interface

	No keyword in Cloud	Keyword in Cloud
Search Box	1.53	1.18
Tag Cloud	2.05	1.31

7.3.3 *Presence of relevant keywords in the tag cloud*

The other questions where use of the tag cloud dominated were Questions 5 and 8. These were more specific questions than 2 and 4, referring to a particular subject or topic:

5. *Paste the title of an article related to biology.*
8. *Paste the title of an article related to NASA.*

With these questions however, there was a relevant keyword present in the tag cloud for the majority of sessions. The keyword *NASA* was present in the tag cloud for every session. The keyword *biology* was present in the tag cloud for seven of the nine sessions. When present

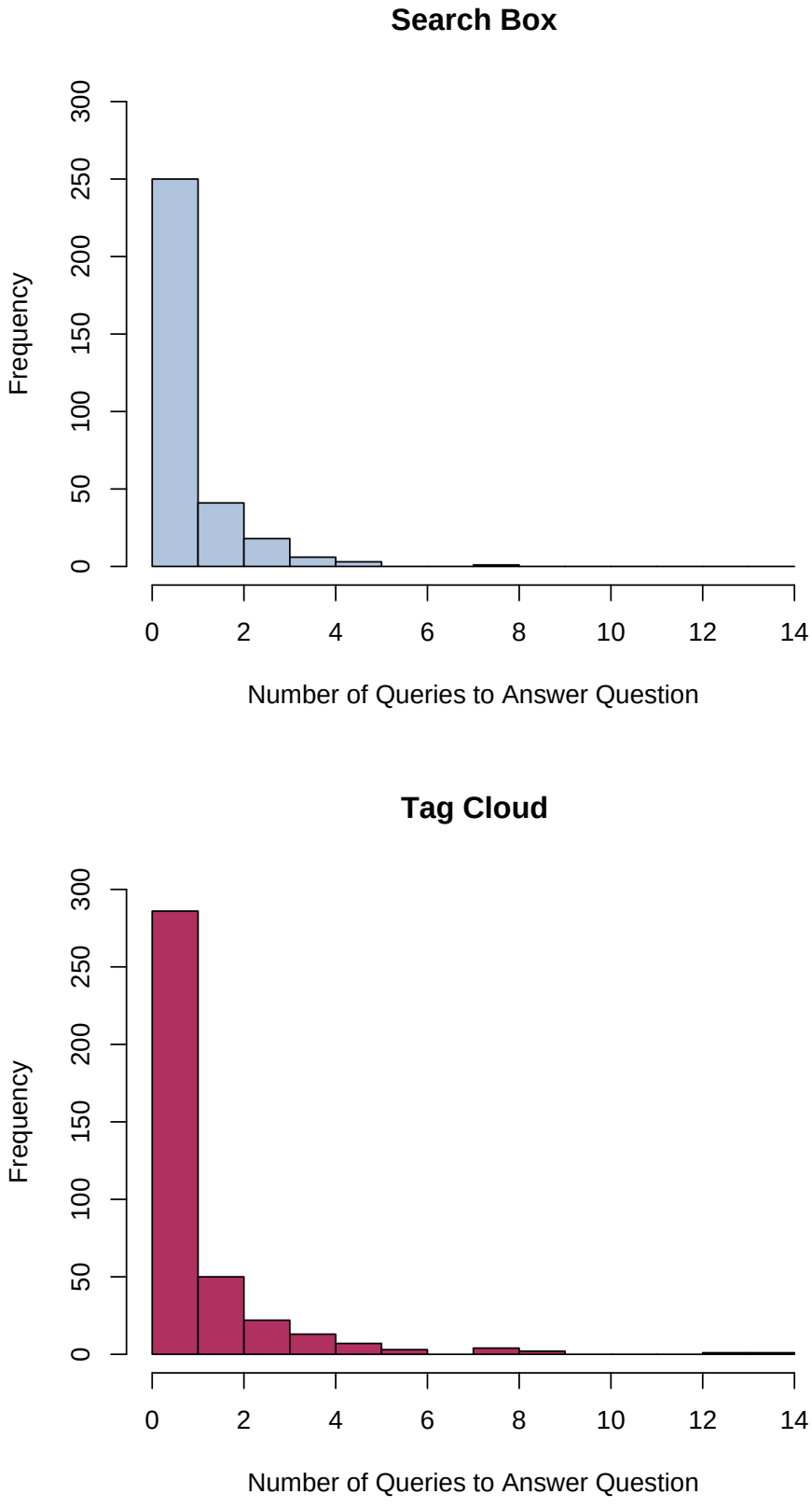


Figure 28: Number of queries required to answer questions where participants relied solely on a single interface

in the tag cloud, both keywords were relatively large in proportion to the surrounding text. A preference for the tag cloud with these questions makes sense, since clicking on a single keyword requires less effort than typing search terms and submitting a form.

Investigating this further, we find that last-strike usage of the tag cloud increased whenever there was a relevant keyword present in the tag cloud (see Table 32).

Table 32: Last-Strike Queries when relevant keywords were present in tag cloud.

Keyword in tag cloud:	No	Yes
Did Not Search:	65 (14.3%)	31 (7.1%)
Search Box:	194 (42.8%)	173 (39.6%)
Tag Cloud:	194 (42.8%)	233 (53.3%)

This is consistent with participant comments that clicking on a single keyword was faster than typing it into the search box. For example:

> *It was quicker, instead of typing in a search word/s, you can just click on the provided keywords to find a relevant article.*
> *[C]licking is fast[er] than typing...*
> *When the question had a tag associated with it, [it] was faster to use the tag*

7.3.4 *Participants' preference*

In the exit survey, a majority of participants stated a preference for the search box (Table 33). Participants' reasons for their preference were coded and tabulated as shown in Table 34, below. There was a wide variety of responses; however, the top three reasons are particularly interesting since they reflect the findings regarding 'browsing versus finding.'

Table 33: Participants' preferences from the exit survey.

SEARCH	DON'T KNOW	TAG	DID NOT ANSWER
35 (39.33%)	11 (12.36%)	26 (29.21%)	17 (19.10%)

A factor to consider here is the makeup of the participant group. As mentioned previously, I recruited participants from a first-year en-

Table 34: Reasons for choosing one interface over another.

RESPONSES	REASON CATEGORY
10	Search box allowed greater specificity
8	Depends on information-seeking task
7	The tag cloud was faster
4	The search box was faster
4	The tag cloud 'suggests' useful query terms
4	The search box was easier to use (or the tag cloud was more difficult)
4	The search box is familiar or the tag cloud unfamiliar
3	The tag cloud was easier to use (or the search box was more difficult)

gineering class. Among such a group there is likely to be a high level of computer literacy and familiarity with information-seeking tools such as search engines. At the same time, very few of the participants appeared to be familiar with the concept of tagging or tag clouds. Only two participants indicated that they used a social bookmarking service, and only ten of the 89 participants indicated that they kept a blog. This may have lead to a greater preference for the search box due to its familiarity. Comments from the exit survey support this, for example:

> *I am used to searching with search boxes on databases and the internet*
> *[I preferred the search box because I am] used to it*
> *[The search box is] a system I am used to*

7.3.5 *Tag cloud as a visual summary*

There is evidence to suggest that formulating a free text search query requires greater mental activity than browsing a list of links [83]. In order to formulate a query, a participant must *recall* relevant words to enter into the search box—a cognitively more expensive task than *recognising* relevant words in a list [101]. A number of the comments from the exit survey indicated that the tag cloud 'suggested' useful search terms, saving the participant some cognitive effort. Examples of participants' comments include:

[T]he tags made it so you didn[']t need to think of what you wanted to find.
The tags invoked ideas immediately in my mind, making access easier
When you want to search something but you don't know where to start, using tags are the right way to start.

In the experiment sessions, I also observed that participants for whom English was a second language seemed to find the tag cloud particularly helpful in this regard. This suggests that tag clouds may have the potential to increase accessibility for people performing information-seeking tasks in a second language. More research is needed to confirm this hypothesis however.

7.3.6 *Tag cloud occlusion*

One disadvantage of the tag cloud is that when used in this manner, it does not allow direct access to all articles in the database. When the system presented search results from a tag based query, it also showed a cloud of co-occurring tags as displayed in Figure 29, below. Clicking on one of these co-occurring tags presented articles tagged with *both* keywords. Thus, if an article had no tags in the front tag cloud (that is, in the 70 most popular tags), then it was inaccessible from the tag cloud interface.

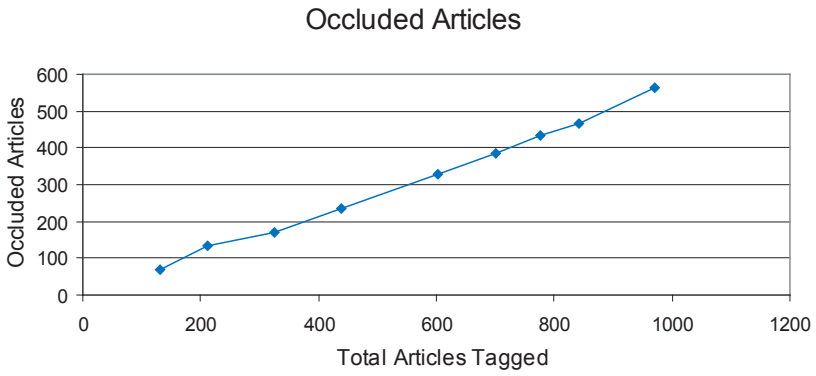


Figure 29: Articles inaccessible from the tag cloud.

In every session of the experiment, a significant number of articles were not accessible from the tag cloud. Interestingly, while the number of tags in the cloud remained constant (70 tags), the number of

occluded articles increased proportionally to the total number of articles (see Figure 29). The percentage of occlusion remained roughly constant at about 55.5% for each session (see Figure 30). This inability to make all articles accessible from a single page is a major disadvantage compared with the search box. Hence, tag clouds are clearly not sufficient as the sole means of navigating a folksonomy dataset when used in this manner.

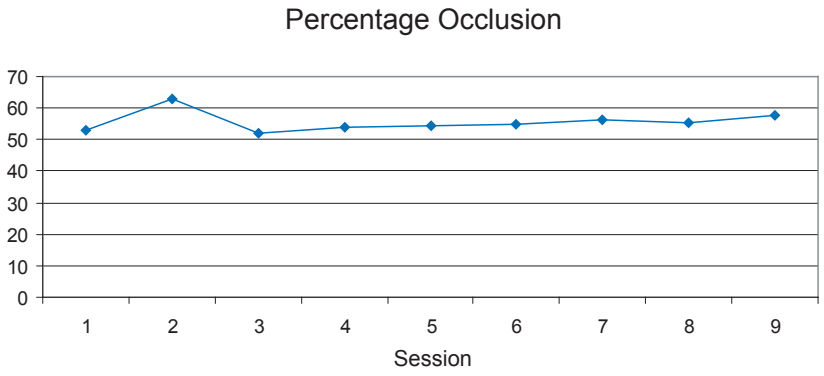


Figure 30: Percentage of articles not accessible from the tag cloud.

7.4 CONCLUSION

This chapter reports on an attempt to answer the question: *Do tag clouds provide value to people seeking information from a folksonomy data set?* After observing that a majority of questions were answered using the tag cloud, we conclude that the tag cloud does indeed provide value. The tag cloud has a number of positive attributes:

- It is particularly useful for browsing or non-specific information discovery. This is partly because tags are best suited to broad categorisation, as opposed to specific characterisation of content. The tag cloud also reduces the cost of a query, since a single click requires less effort than typing search terms and submitting a form.
- The tag cloud provides a visual summary of the contents of the database. A number of participants commented that this gave them an idea of where to begin their information seeking. The tag cloud provides potential for a two-phase search process where: 1) browsing activity using the tag cloud familiarises the

information seeker with the domain, and 2) a search mechanism allows more focussed queries based on the information gained from browsing.

- It appears that scanning the tag cloud requires less cognitive load than formulating specific query terms. That is, scanning and clicking is 'easier' than thinking of appropriate query terms then typing them into the search box. I hypothesise that this may be particularly useful for people seeking information in a second language.

The tag cloud is not without its limitations however:

- A corollary of the tag cloud's suitability for general browsing is its unsuitability for seeking specific information. Many of the participants commented that the tag cloud did not allow them to narrow their search sufficiently to answer the given questions.

- For this experiment the results indicated that roughly half the articles in the dataset were inaccessible from the tag cloud. Most tagging systems mitigate this by including tag links when viewing individual articles, thus exposing some of the less popular tags. However, this is not necessarily useful when someone is seeking specific information.

In light of these disadvantages, it is clear that the tag cloud is not sufficient as the sole means of navigating a folksonomy dataset. Some form of search facility or supplementary classification scheme is necessary to expose all the articles. On the other hand, it would be unfair to say that the tag cloud offers no value as a navigation aid. Where relevant keywords are present, clicking on a tag is quicker and easier than entering query terms into a search box. It also provides a useful visual summary of the contents of the database.

While this study shows that tag clouds are a useful method for supporting resource discovery, it does not address the other drawbacks of folksonomies; there remains the problems of polysemy and synonymy, and how to scale down folksonomies for smaller systems. Some of these issues are addressed in the next chapter through the use of clustering algorithms.

8

COMPARISON OF FOLKSONOMY CLUSTERING TECHNIQUES

Chapter 7 showed that tag clouds are a useful method of presenting folksonomy data in conjunction with other information retrieval tools. The tag cloud provides a useful visual summary of the content in the information system, and is easier to use than full-text search. Tag clouds, however, do little to address some of the other shortcomings of the folksonomies such as homonymy, synonymy and tag noise. One common method proposed to address these shortcomings is the use of clustering techniques. Many authors and practitioners have implemented various means of clustering folksonomies. However, each approach is different and utilises different clustering techniques. This begs the question: *which clustering technique is best suited to clustering folksonomies?*

To address this question we implemented four different clustering algorithms and applied them to a folksonomy dataset. Two of the algorithms were well-known clustering algorithms that have been applied to folksonomies, while the other two algorithms were from the transactional data clustering literature. Of the four algorithms tested, we found that the ROCK algorithm performed best, but was also the most algorithmically complex.

8.1 INTRODUCTION

In previous chapters we have proposed folksonomies as a useful tool for organising information where a dedicated administrator is not available. Folksonomies are attractive because they generate collaboration, reflect the vocabulary of participants, and are well suited to multimedia data. Proponents of folksonomies characterise them as open, democratic and flexible, as opposed to exclusive, rigid, hierarchical taxonomies. For all their advantages however, folksonomies are not without their drawbacks.

Folksonomies have a number of drawbacks, or disadvantages. Firstly, there are the well known issues of synonymy, homonymy and tag noise; secondly, some tags will have different meanings when applied to certain items; and thirdly, some people will apply tags which have personal meaning for them, but are of little semantic value. These issues tend to reduce the effectiveness of information retrieval in a

folksonomy.

In many cases, practitioners and authors have used clustering techniques to address the shortcomings of folksonomies. By clustering together resources with common tags, there is potential to identify synonyms and differentiate homonyms. While many authors have proposed ways of using clustering techniques with folksonomies, each approach is different. In the literature on folksonomy clustering, very few authors justify their choice of algorithm and simply report that their approach appears to work. This is surprising however, given the bewildering array of clustering algorithms available.

This gap in the folksonomy clustering literature led us to compare different algorithms for clustering folksonomies. We implemented four algorithms: two based on well-known clustering techniques that authors have previously applied to folksonomies, and two algorithms developed for clustering market-basket data. We compared the clusters produced by each algorithm with a human-generated category scheme to evaluate their performance. We found that one of the market-basket algorithms,ROCK, produced the best results, but is algorithmically very complex and does not scale well.

To compare the clustering results I propose a new method for comparing different clustering solutions. This new method allows a clustering solution to be compared with a human-generated classification scheme, even if the scheme is hierarchical (having multiple levels) or category membership is not mutually exclusive. Unlike other cluster validity measures, this method does not have a strong bias towards very small or very large cluster sizes.

8.2 BACKGROUND

Applying clustering techniques is one approach used to improve information retrieval in folksonomies. A clustering algorithm attempts to organise things (or mathematical representations of things) into groups. It does this without supervision—that is, no information is available on the 'correct' grouping; a correct grouping may not even exist. Putting things into groups is useful as it helps people cope with large amounts of information by allowing us to exploit similarities.

Clustering is a very broad field arising out of many different fields of study, including image processing, information retrieval, biology, psychiatry, psychology, archaeology, geology, geography and marketing [64]. It would be impossible to give a comprehensive survey of clustering techniques and their application in this chapter. Jain et al.

[64] provide a good introduction to the subject however, and make some helpful distinctions between different kinds of clustering methods.

8.2.1 *Why Cluster Folksonomies?*

Before describing how clustering techniques operate, it is useful to ask why we should use them at all. What advantage is there in organising folksonomy data into groups? We propose three reasons why one might wish to use clustering with folksonomies:

1. Clustering may help discover groups of people sharing common interests. This kind of functionality could serve many purposes, particularly in an enterprise environment. People may discover that they share common interests with someone whom they would not normally associate with their field. Communities of practice may be discovered by examining resources people have in common.

2. Clustering can generate an automated or semi-automated taxonomy/hierarchy. While the 'flat' nature of folksonomies is part of their appeal, hierarchies can be very useful for dealing with complex information. Clustering techniques can be used to create hierarchical navigation schemes to help people navigate a tagged collection.

3. Clustering can help disambiguate tags. Often, resources in a tagging system have multiple tags with synonymous meanings. By analysing co-occuring tags, we can group together synonymous tags, while homonymous meanings for a tag can be separated. One early adopter of this approach is the Flickr* photo sharing service. Given a popular tag, Flickr clusters related photographs into groups. For example, looking at the clusters for the word 'tiger', we can see that they fall neatly into different semantic categories: big cats, a version of the Macintosh operating system, the Tiger Swallowtail Butterfly, and Tiger Lilies.

8.2.2 *Clustering Techniques*

In general, a folksonomy dataset consists of three distinct entities: *people*, *tags* and *resources*. Clustering may be performed on any one

* http://www.flickr.com

of these entities. People may be clustered according to the tags they use; resources may be clustered according to the tags assigned them; and the tags themselves may be clustered according to the tags with which they co-occur. For the purposes of this article, we focus on the clustering of resources, based on tags they share in common.

Before applying a clustering technique, we must first transform the problem into a numeric representation suitable for a clustering algorithm. The most common way of numerically representing a resource is to count the frequency of each tag assigned to it. This is a similar to the word frequency counts used in document clustering. In the following sections, we will introduce notation to describe a folksonomy dataset and outline the similarity measures used to group resources together.

Notation

A folksonomy dataset can be represented mathematically in a number of different ways. Here, we use both a vector notation and a set notation. In the vector notation, a folksonomy $\mathbf{F}$ is represented by an $n \times d$ matrix:

$$\mathbf{F} = \begin{pmatrix} f_{1,1} & \dots & f_{1,d} \\ \vdots & \ddots & \vdots \\ f_{n,1} & \dots & f_{n,d} \end{pmatrix}$$

where each element, $f_{i,j}$, represents the frequency of tag j in resource i. The dimensionality, d, is the number of unique tags in the folksonomy. A single row in $\mathbf{F}$ forms a vector of length d that represents a single resource. That is:

$$\mathbf{r}_i = (f_{i,1}, \dots f_{i,d})$$

In set notation, the folksonomy F is represented by a set of resources:

$$F = \{r_1, \dots r_n\}$$

where each resource r, represents a set of tags:

$$r = \{t_1, \dots t_m\}$$

A clustering solution λ is represented as a set of clusters:

$$\lambda = \{C_1, \dots C_k\}$$

Where each cluster C_i is a set of resources.

Similarity Measures

Many (though not all) clustering algorithms work by calculating similarity between resources. The algorithm then aims to maximise the similarity within each cluster and minimise the similarity between different clusters. Two common similarity measures are cosine similarity and the Jaccard similarity.

Conceptually, the *cosine similarity measure* corresponds to the angle between two vectors. We calculate cosine similarity by taking the dot product of two vectors and normalising by their length.

$$\begin{aligned} \text{sim}(\mathbf{r}_i, \mathbf{r}_j) &= \frac{\mathbf{r}_i \bullet \mathbf{r}_j}{|\mathbf{r}_i|\,|\mathbf{r}_j|} \\ &= \frac{\sum_{k=1}^{d} f_{i,k} f_{j,k}}{\sqrt{\sum_{k=1}^{d} f_{i,k}^2}\sqrt{\sum_{k=1}^{d} f_{j,k}^2}} \end{aligned} \tag{8.1}$$

The *Jaccard similarity* is the number of tags two items have in common divided by the total number of tags for the two items. Using set notation, the similarity equation is:

$$\text{sim}(r_i, r_j) = \frac{|r_i \cap r_j|}{|r_i \cup r_j|} \tag{8.2}$$

This equation works well when the frequency of each tag is either one or zero; however, it does not deal very well with tags appearing more than once. To address this, Strehl and Ghosh [130] propose an *extended Jaccard similarity*. Using a vector notation, the equation is:

$$\text{sim}(\mathbf{r}_i, \mathbf{r}_j) = \frac{\mathbf{r}_i \bullet \mathbf{r}_j}{|\mathbf{r}_i|^2 + |\mathbf{r}_j|^2 - \mathbf{r}_i \bullet \mathbf{r}_j} \tag{8.3}$$

If no tags occur in a resource more than once, these two equations are equivalent.

Other similarity measures are possible, but these two are the main measures used with the clustering algorithms we examined.

8.3 PREVIOUS WORK

As indicated previously, a number of authors have begun to mention the application of clustering techniques to folksonomy based data. In this section, we examine how these techniques have been applied and the purpose of their application.

Brooks and Montanez [18] examined the folksonomy generated by the Technorati tagging service. They 'show that clustering algorithms can be used to reconstruct a topical hierarchy among tags, and suggest that these approaches may be used to address some of the weaknesses in current tagging systems'. In particular, they use a hierarchical agglomerative algorithm with a cosine similarity measure, citing Cutting et al. [28]. This citation is surprising since Cutting et al. argue that hierarchical agglomerative algorithms are extremely limited due to their $O(n^2)$ running time and a tendency to 'undesirable chaining behaviour, typically forming elongated straggly clusters.' Nevertheless, they report positive results from their experiment, claiming that their generated hierarchy largely 'mirrors that seen in hand-built taxonomies such as those built by dmoz.org or Yahoo!.'

Of special note regarding the Brooks and Montanez algorithm is that, although it uses tags to label clusters, tags are not actually involved in computing similarity measures. Instead, the algorithm uses word frequencies from the blog articles. After performing reporting on some initial experiments the paper argues that 'tags are useful for grouping articles into broad categories, but less effective in indicating the particular content of an article'. This may well be an effective approach when textual data is available for tagged resources (such as the blog articles they clustered), but tagged resources are not always textual. Services such as Flickr, Last.fm† and PodNova‡ tag multimedia resources are not amenable to this approach. Hence, in our experiments we focus on tags as the sole source of information for clustering.

Unlike Brooks and Montanez [18], Begelman et al. [13] developed their own clustering algorithm for folksonomies based on patterns they observed in the frequency distribution of tags. They observed that in a broad folksonomy with many taggers, the frequency distribution for co-occurring tags tends to display a kink (Figure 31). Assuming that the tags occurring to the left of the kink are more strongly related than those on the right are, they develop a heuristic

† http://www.last.fm/
‡ http://www.podnova.com/

for finding this point.

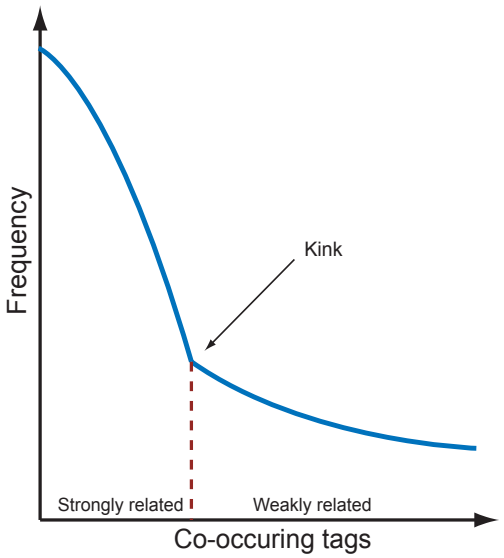


Figure 31: Observed distributions of co-occurring tags

Using this heuristic, Begelman et al. [13] group strongly related tags together. The groups form the basis of a weighted undirected graph, which is then clustered using a spectral clustering algorithm. Using this clustered graph, the authors propose a method for using the clustering results to present semantically related tags as an aid to browsing a folksonomy.

In this case, Begelman et al. cluster tags rather than resources, since their purpose is to present potentially relevant tags to a user. That is, instead of grouping resources that share common tags, they group tags that occur together frequently. While this approach suits the aims for their particular system, it is limited in its suitability for broader application.

Another contribution to this area is the work of Lambiotte and Ausloos [73, 74], even though they do not use the word 'clustering' in their papers. Rather, they discuss methods of representing folksonomies as graphs, and outline 'percolation techniques that consist in removing less correlated links and observing the shaping of disconnected islands' [73]. The result however, is essentially a form of clustering.

The approach of Lambiotte and Ausloos is the most general, since it allows for different representations of the folksonomy according to what is of interest. For example, the method can group people according to common resources, or people according to common tags, or resources according to common people, and so on. This allows

them to investigate the structure of the folksonomy in a number of different ways.

Lambiotte and Ausloos form their graph representation by using a cosine similarity measure to determine the weight of links between nodes. A weight of zero means there is no link between nodes. The nodes can be people, resources, or tags, and the similarity measure can be calculated using common tags, common resources, or common people. By removing links with weights below a certain threshold, 'islands' in the graph appear which form the clusters.

Lambiotte and Ausloos use this representation as a way of investigating *diversity* in a folksonomy. That is, by examining the tags in a person's collection, they develop a way of measuring how broad the resources in that collection are. For example, the tagging service Last.fm allows people to tag music that they listen to with genres. Diversity would give a measure of how broad or narrow is that person's taste in music.

Because Lambiotte and Ausloos are proposing more of an analysis method than a practical application of clustering techniques, we have not included their method in our experiments. Their approach to modelling the folksonomy however, is extremely powerful and deserves recognition.

Similarly to Lambiotte and Ausloos, Schmitz [116] does not define a clustering method as such. Instead, Schmitz attempts to induce a faceted ontology from the Flickr tag set using a subsumption-based model. The result is a group of tree structures containing parent tags, and tags considered sub-topics of the parent tag.

The subsumption model looks at co-occurrences of tags. Where two tags occur together frequently, it attempts to determine if one is a subconcept of the other. They report initial results from a simple model, but state that their work on a full probabilistic model is still ongoing.

Another work mentioning tag clustering is that of Bielenberg and Zacher [14] in which they describe their use of a hierarchical algorithm to form clusters. The results of the clustering algorithm form the basis of a tag cloud for each cluster[§]. They argue that the use of a tag cloud in this manner gives a person using the system 'a quick impression of the structure of her archive or individual context.'

In their clustering algorithm, Bielenberg and Zacher use an unusual similarity measure that appears to be a variation on the Jaccard similarity. It is calculated by counting the number of tags both items have in common, divided by the number of tags the items do not have in

§ see Chapter 7 for a discussion of tag clouds

common:

$$\text{sim}(r_1, r_2) = \frac{|r_1 \cap r_2|}{|r_1 \cup r_2| - |r_1 \cap r_2|}$$

While most similarity measures tend to have a range of [0,1], the value of this similarity measure is infinity if both objects have identical tags.

Aside from the similarity measure, Bielenberg and Zacher's work gives very little detail regarding their implementation. They write that describing hierarchical clustering in detail is beyond the scope of their paper, and refer the reader to Jain et al. [64] for an overview. The paper leaves the reader to assume that it worked well enough to suit their purposes.

Most of the works cited above appear to give little justification for their choice of clustering algorithms. Why did they choose the particular algorithm they did? What makes it most suitable for clustering folksonomies? Those that develop novel techniques fail to explain why they reject existing techniques. Given the bewildering array of clustering techniques available, which are best suited to the properties of a folksonomy data set? In order to address this question, we conducted our own experiments comparing the performance of various clustering techniques.

Clustering techniques have been discussed in detail in the document clustering literature. However, the tag frequencies in a folksonomy are different to the word counts obtained from written documents. Folksonomies do not have the same need to screen stop-words, since we assume that people tagging will not include words such as 'to', 'of', or 'and', without good reason. Folksonomies tend to exhibit lower dimensionality than written document analysis, but folksonomies will tend to be sparser. When two resources in a folksonomy share a tag, it is more likely that they are related, than if two documents share a single word. In short, we cannot assume that what works best for document clustering will work equally well with folksonomies.

8.4 METHOD

Our aim in this experiment was to begin answering the question: *Which clustering techniques are best suited to clustering folksonomies?* To do this, we implemented four clustering algorithms, and applied them to a folksonomy dataset. We then compared the clusters to a human-

generated categorisation scheme to evaluate the quality of the results. We found that all of the algorithms tested produced reasonable results, showing significant improvement over random assignment. Of the four techniques, the ROCK algorithm produced the results most similar to the human-generated categorisation. However, the ROCK algorithm is also the most computationally complex of the techniques tested.

8.4.1 *External Cluster Quality Measures*

The question 'Which clustering techniques are best suited to clustering folksonomies?' begs another obvious question: 'How can we compare one clustering with another?' Since clustering is by nature an unsupervised technique, comparing the 'goodness' of clusters is a difficult task. Often, a researcher will develop a new clustering algorithm by first defining a criterion function—that is, a measure of what a 'good' clustering looks like—then create an algorithm to maximise that function. Naturally, if we use this criterion function to measure the results of the clustering, then this algorithm should out-perform most other algorithms, since it maximises that criterion by design.

To avoid this issue, we compare the results of the clustering algorithm to a human-created classification scheme. We call these comparisons external measures of cluster quality. In the document clustering and data mining literature, authors have proposed a number of external measures. These include purity, entropy, and the F-measure [129, 130, 146]. After attempting to evaluate our results using these measures however, we found that these approaches were not appropriate to the non-exclusive categorisation provided by our dataset.

To explain the external quality measures, we must first introduce some notation (summarised in Table 35). We compare the clustering result λ with a categorisation scheme κ containing g categories. For this clustering solution and categorisation scheme, let n_{ij} be the number of objects in cluster j belonging to category i, and let n_j be the number of items in cluster C_j. Similarly, let n_i be the number of items in the dataset belonging to category i.

Purity

The purity measure finds the most dominant category in each cluster, and then counts the percentages of items belonging to that category. Thus, it gives a measure of how well each cluster corresponds to a cat-

Table 35: Notation for external cluster quality measures

λ	A clustering solution.
k	The number of clusters in clustering solution λ.
κ	Categorisation scheme for comparison.
g	the number of categories in categorisation scheme κ.
n_{ij}	The number of items in cluster j belonging to category i.
n_j	The number of items in cluster j.
n_i	The number of items in category i.

egory in the dataset. Mathematically, purity can be defined as follows [130]:

$$\begin{aligned} P(\lambda) &= \sum_{j=1}^{k} \frac{1}{n_j} \max(n_{ij}) \\ &= \frac{1}{n} \sum_{j=1}^{k} \max(n_{ij}) \end{aligned}$$

The disadvantage of the purity measure is that it takes no account of inter-cluster scatter. That is, there is no penalty if items from a single category are divided across many clusters. A clustering solution which places every item in its own cluster will always give maximum purity. Thus we found that purity was not a suitable measure for our experiments.

Entropy

Unlike the purity measure, which measures each cluster against a dominant category, entropy takes each category in the cluster into account. If many different categories are present in a single cluster, then the cluster will receive a higher entropy score. Conversely, if a cluster contains mainly items belonging to a single category, then it will have a low entropy score. Strehl and Ghosh [130] define the entropy for a single cluster C_j as:

$$E(C_j) = \frac{1}{\log(g)} \sum_{i=1}^{g} \frac{n_{ij}}{n_j} \log\left(\frac{n_{ij}}{n_j}\right)$$

Here, lower entropy scores are better. Following Steinbach et al. [129], the total entropy for a cluster solution is the sum of entropies for each cluster, weighted by the size of the cluster:

$$\begin{aligned} E(\lambda) &= \sum_{j=1}^{k} \frac{n_j E(C_j)}{n} \\ &= \frac{1}{n \log(g)} \sum_{j=1}^{k} \sum_{i=1}^{g} n_{ij} \log\left(\frac{n_{ij}}{n_j}\right) \end{aligned}$$

One difficulty with the entropy measure is that it is strongly weighted towards smaller clusters. That is, a single large cluster will usually receive a much poorer score than a group of smaller clusters. This is especially the case if the category set κ is not mutually exclusive. Thus, a clustering solution which forms smaller, evenly-sized clusters will perform better than a solution consisting of varying size clusters, even if the algorithm is identifying strong similarities. Weighting the overall entropy score by cluster size exaggerates this issue even further. Hence, we found that in practice the entropy measure ranked our results in reverse order of actual quality, since the number of items in each category was not uniform.

Mutual Information

Strehl and Ghosh [130] propose that the mutual information measure is a better approach as it is not so biased towards smaller clusters. We define mutual information as follows:

$$\begin{aligned} M(\lambda) &= \sum_{i=1}^{g} \sum_{j=1}^{k} \frac{n_{ij}}{n_j} \log\left(\frac{\frac{n_{ij}}{n_j}}{\frac{n_i}{n}\frac{n_j}{n}}\right) \\ &= \sum_{i=1}^{g} \sum_{j=1}^{k} \frac{n_{ij}}{n_j} \log\left(\frac{n_{ij} n}{n_i n_j^2}\right) \end{aligned}$$

Strehl and Ghosh argue that mutual information considers higher order dependencies. In practice, I found that mutual information does rank algorithms in the correct order (unlike the entropy measure); however, the measure still gives a maximum value when each item is in its own cluster.

F-Measure

In their comparison of document clustering techniques, Steinbach et al. [129] use the F-measure proposed by Larsen and Aone [75] to evaluate cluster quality. The F-measure combines the ideas of precision and recall from information retrieval. For a cluster C_j and category i:

$$\begin{aligned} \text{Recall}(i,j) &= \frac{n_{ij}}{n_i} \\ \text{Precision}(i,j) &= \frac{n_{ij}}{n_j} \end{aligned}$$

The F measure is given by:

$$F(i,j) = \frac{\text{Recall}(i,j)\,\text{Precision}(i,j)}{\text{Recall}(i,j) + \text{Precision}(i,j)}$$

Which simplifies to:

$$F(i,j) = \frac{2n_{ij}}{n_i + n_j}$$

The overall F measure for the document set is given by:

$$F(\lambda) = \frac{1}{n} \sum_{i=1}^{k} n_i \max\left(F(i,j)\right)$$

where the maximum is taken over all clusters in the solution λ.

In practice, I found that the F-measure produced results that ranked the clustering algorithms in the correct order. However, the F-measure is maximised when cluster sizes are largest, that is, when all items belonged to a single cluster. Like purity, this approach assumes that each cluster should represent a single category.

The problem with all these measures is that they are not well suited to the categorisation schemes employed by the two datasets I used for comparison. That is, because an article can belong to multiple categories, it is difficult to determine if a cluster 'correctly' identifies a category. Thus, I developed a novel measure to compare the clustering results of the two datasets.

8.4.2 *My Approach*

To form a 'good' cluster, items within a cluster should be as similar to each other as possible. At the same time, we wish to keep items belonging to the same category together. Thus, we propose a combination of two measures: intra-cluster similarity (ICS) and category scatter (CS).

Intra-Cluster Similarity

We measure ICS for the entire clustering solution by taking the weighted mean pair-wise similarity between clusters:

$$\mathrm{ICS}(\lambda, \kappa) = \sum_{j=0}^{k} \frac{n_j}{n} \mathrm{ICS}(C_j)$$

where the pair-wise ICS for a single cluster is given by:

$$\begin{aligned} \mathrm{ICS}(C_j) &= \frac{1}{n_j P_2} \sum_{a,b \in C,\, a \neq b} \mathrm{sim}(a,b) \\ &= \frac{1}{n_j(n_j-1)} \sum_{a,b \in C,\, a \neq b} \mathrm{sim}(a,b) \end{aligned}$$

and $\mathrm{sim}(a,b)$ is the Jaccard similarity (Equation 8.2), calculated according to the categorisation scheme κ. That is, $\mathrm{sim}(a,b)$ is calculated using categories rather than tags.

ICS is similar to the entropy measure; the more diversity within the cluster, the lower the ICS will be. In the case where each item is in its own cluster, ICS will be one.

Cluster Scatter

We also wish to measure how spread out each category is across the cluster. Ideally, all the items from a single category will be contained within a single cluster. We define s_i as the number of clusters that contain items in a given category i. That is, the number of clusters that category i has been scattered across. We then normalise this value to give the scatter measure for an individual category:

$$\mathrm{Scatter}(i) = \frac{s_i - 1}{n_i - 1}$$

If every item in a category is placed in a separate cluster then scatter will be one.

To calculate the scatter across an entire clustering solution λ we take a weighted sum of scatter for each category:

$$\text{Scatter}(\lambda, \kappa) = \frac{\sum_{i=1}^{g} n_i \text{Scatter}(i)}{\sum_{i=1}^{g} n_i}$$

It is important to note that, because an item may belong to multiple categories, $\sum n_i \neq n$.

The Q Measure

The ICS and CS measures take into account different aspects of what makes a 'good' clustering. A good clustering solution will have a high ICS, but a low CS. Hence, we wish to define a quality measure that will combine both these aspects. To do this, we specify a number of criteria:

1. When both factors are optimal, the quality measure is maximised.
2. When every item is in its own cluster, the quality measure should be minimised, since this clustering adds no value.
3. Similarly, a clustering solution with one cluster containing every item should have a minimum value, since this clustering adds no value either.

As a first attempt, we combined the terms using the formula:

$$Q(\lambda, \kappa) = \frac{2\text{ICS}(\lambda, \kappa)\,(1 - \text{CS}(\lambda, \kappa))}{\text{ICS}(\lambda, \kappa) + (1 - \text{CS}(\lambda, \kappa))}$$

This fulfils the first and second criteria set out above, however, it fails to meet the third criterion. This is because the dataset has some self-similarity, even when every item is in a single cluster. Hence, we scale the ICS by this residual similarity. We label the scaled value ICS′. That is:

$$\text{ICS}'(\lambda, \kappa) = \frac{\text{ICS}(\lambda, \kappa) - \text{ICS}(0)}{1 - \text{ICS}(0)}$$

where ICS(0) is the residual similarity. This gives us the equation for our Q measure:

$$Q(\lambda, \kappa) = \frac{2ICS'(\lambda, \kappa)\,(1 - CS(\lambda, \kappa))}{ICS'(\lambda, \kappa) + (1 - CS(\lambda, \kappa))}$$

This measure fulfils all three of our criteria.

8.4.3 *Choice of Datasets*

In this thesis, we are particularly interested in the use of folksonomies in less-than-ideal situations, such as might be found in a SME. To reflect this, we deliberately chose a small, noisy dataset collected as part of other research into folksonomies. However, such a dataset is not necessarily typical of folksonomies. So, we also tested the algorithms on a larger, more typical folksonomy dataset from a popular social bookmarking service.

The first dataset we used consisted of 973 items tagged by 89 participants as described in Chapter 7. The items consisted of articles from the Slashdot news website that have been categorised (presumably) by humans to allow Slashdot users to subscribe to different topics of interest. We chose this categorisation scheme as the basis for our external comparison with clustering results.

The design of the Slashdot service, however, means that one article may fall into multiple categories such as science, technology, humour, etc. That is, the Slashdot articles are *categorised*, rather than *classified*.

Appendix C shows the number of articles in each of the Slashdot categories. Appendix D shows the frequency of each tag in the dataset that occurred more than once. By comparing the larger categories with the more frequent tags, we can see that there are commonalities (see Table 36 for a summary). The top three Slashdot categories are *Science*, *Space* and *Technology*, while the most frequent tags were *NASA*, *Space* and *Science*. It is interesting to note that the tags identified some popular themes in the dataset that are not identified as categories, namely, *global warming* (and also *Greenhouse*) and *health*.

The second dataset examined is taken from the popular social bookmarking service *RawSugar*[¶] through a publicly available XML interface. We compared the data from this site with human generated categorisations from the Open Directory Project (DMOZ)[‖].

¶ http://www.rawsugar.com
‖ http://dmoz.org/

Table 36: Top ten tags and categories from the Slashdot dataset.

CATEGORIES		TAGS	
Science	473	NASA	68
Space	196	Space	62
Technology	113	Science	53
News	81	Mars	18
Biotech	79	biology	14
NASA	42	global warming	13
Power	25	health	13
Politics	23	Moon	13
Robotics	19	physics	12
Education	18	Greenhouse	11

Due to the large size of these two data sources, comparing the entire corpus of both would have been impractical. Instead, I selected all the Uniform Resource Locators (URLs) from the DMOZ data in the category *Computers>Programming*. An automated script then downloaded tags for any URLs in RawSugar that also occurred in the DMOZ set. This process resulted in a set of 1749 URLs, with 2052 unique tags.

To reduce processing, both datasets were *trimmed* to remove any items from the set which had no tags in common with any other item. After trimming the Slashdot dataset contained a total of 616 articles, while the RawSugar dataset contained 1707 URLs. The top ten tags and categories are summarised in Table 37.

Table 37: Top ten tags and categories from the RawSugar dataset.

TAGS		CATEGORIES	
programming	694	Languages	14281
development	387	Languages>Java	2854
software	279	Languages>PHP	1458
java	229	Component_Frameworks	1342
web	210	Internet	1305
system:imported	196	Languages>PHP>Scripts	1048
php	184	Languages>Perl	908
tools	149	Languages>Fortran	893
opensource	148	Languages>C++	789
reference	139	Methodologies	583

8.4.4 *Implementation of Algorithms*

Given the extensive number of clustering techniques available, it would be impractical to compare them all. Hence, we chose two popular algorithms, K-means and Hierarchical Agglomerative, and compared these with two algorithms designed to cluster transactional (market basket) data, as folksonomies appear to share a number of similarities with this kind of data.

Random Baseline

In order to compare the results of the clustering algorithms we measured a baseline by assigning articles to clusters at random. We did this in a two-stage process:

1. Create k empty clusters, and then assign one item selected at random to each.
2. Assign each of the remaining items to a cluster selected at random.

This allowed us to achieve the desired number of clusters, but without forcing a uniform cluster size on the result.

We repeated this process, calculating ICS and CS, fifteen times for each value of k.

K-means Clustering

K-means is perhaps the best-known clustering algorithm, as it is reasonably simple to explain and understand. It assumes that the desired number of clusters is known in advance and goes through the following steps:

1. Select k points to be centroids of the clusters. Usually this is done by selecting k resources at random; however, there are many different ways of choosing centroids.
2. Assign each point to the closest centroid, by calculating the similarity between each centroid and each resource.
3. Re-calculate the centroid to be the centre-point of the cluster. For a centroid $\mathbf{c}$ with elements $(c_1, \ldots c_d)$, we calculate the value of each element using: $c_j = \frac{1}{n} \sum_{i=1}^{n} f_{i,j}$, where $f_{i,j}$ are elements of $\mathbf{F}$.

4. Repeat steps 2 and 3 until the centroids do not move.

One advantage of the K-means algorithm is that the running time is rectangular, that is, $O(kn)$. The algorithm depends linearly on the number of items and the desired number of clusters.

Our implementation of the K-means algorithm followed much the same pattern as described above. To select the initial centroids, however, we did not choose items at random. Instead, we created centroids based on the k most frequent tags in the dataset.

To create centroid i, we took the i^{th} most frequent item in the dataset, set the frequency of that item to one, and zeroed the rest of the vector. In our experiments, we found that this approach provided consistently better results than random selection.

Hierarchical Agglomerative Clustering

This is another well-known clustering algorithm, often applied in document clustering applications. The basic steps are as follows:

1. Treat each resource as a single-item cluster. Compute a similarity matrix containing the similarity between each pair of clusters.

2. Using the similarity matrix, find the most similar pair of clusters. Merge these clusters together to form a single cluster. Update the similarity matrix to reflect this change.

Repeat step 2 until no clusters are left or the desired number of clusters is reached.

One drawback of the hierarchical agglomerative algorithm is that they have quadratic running times $O(n^2)$ [28]. Thus, if the number of items to be clustered becomes large, the algorithm takes a long time to complete.

Our implementation of the hierarchical agglomerative algorithm followed much the same pattern as described above. For the purposes of this experiment, we did not construct an actual hierarchy using the hierarchical agglomerative algorithm. Instead, we allowed the algorithm to iterate until it reached the desired number of clusters, and calculated the appropriate external measures.

When deciding how to compare the similarity of two clusters (as opposed to two resources), there are a number of options available. Here, we used the mean cosine similarity between all pairs of resources. That is:

$$\mathrm{sim}(C_a, C_b) = \frac{1}{n_a n_b} \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} \mathrm{cosine}(\mathbf{r}_{a,i}, \mathbf{r}_{b,j})$$

Steinbach et al. [129] show that this method of comparing clusters is superior to other methods across a variety of datasets.

ROCK

Guha et al. [49] developed the ROCK clustering algorithm to cluster items with categorical attributes or market-basket data. They base their algorithm on the number of 'links' between items. The algorithm considers items as linked if they have a Jaccard similarity greater than a certain threshold. They claim that whereas most algorithms consider 'local properties involving only the two points in question, the link concept incorporates global information about the other points in the neighbourhood of the two points'.

At its core, ROCK is an adaptation of the hierarchical clustering algorithm discussed above. Instead of calculating a similarity measure to determine if clusters should be merged, however, the algorithm uses a 'goodness' measure:

$$g(C_i, C_j) = \frac{\mathrm{link}(C_i, C_j)}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}}$$

Here, $\mathrm{link}(C_i, C_j)$ is the number of links between clusters C_i and C_j, n_i is the number of items in cluster C_i, and $f(\theta)$ is a function of the similarity threshold. This is usually something like $\frac{1-\theta}{1+\theta}$, where θ is the similarity threshold.

One disadvantage of the ROCK algorithm is its computational complexity. ROCK is based on the hierarchical agglomerative algorithm, which has an $O(n^2)$ complexity to begin with. The computation of neighbours however, adds to the complexity.

We chose to examine the ROCK algorithm because other authors have used it as a benchmark of what constitutes 'good' clustering results [e. g. 142].

Our implementation of the ROCK algorithm used the data structures suggested by Guha et al. [49]. We used the Jaccard similarity measure (Equation 8.2), for calculating similarities and tested the algorithm for a number of values of θ. The value of θ determines the threshold above which the algorithm considers items to be 'neigh-

bours'. We also defined $f(\theta)$ to be $\frac{1-\theta}{1+\theta}$, as suggested by Guha et al.'s more detailed paper [48].

CLOPE

Yang et al. [146] developed CLOPE to cluster large transactional datasets effectively and quickly. Rather than developing a new similarity function, the CLOPE algorithm attempts to maximise the height-to-width ratio of the cluster histogram. A cluster histogram simply shows the number of times each tag occurs in descending order. For example, if a cluster contained one item tagged (a, b, c), one item tagged (a, c), and one item tagged (a), the histogram would look like that shown in Figure 32 below. For the CLOPE algorithm, height refers to the average height across each tag, while width refers to the total number of unique tags in the cluster.

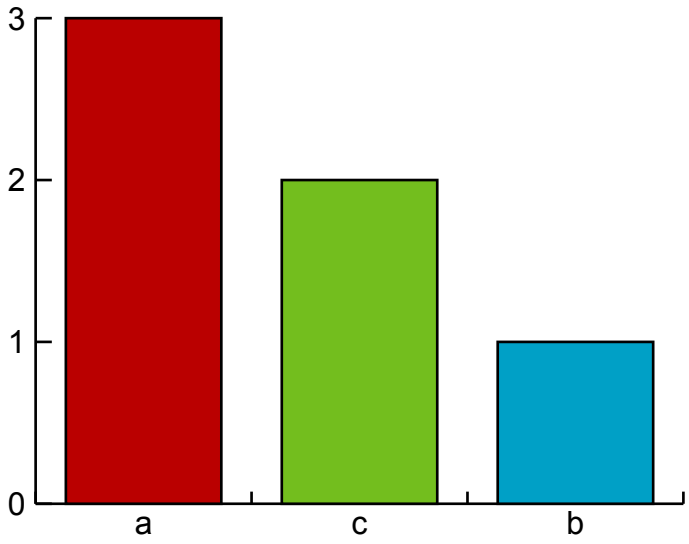


Figure 32: An example of a cluster histogram.

One advantage of the CLOPE algorithm is that it is computationally less expensive than other clustering techniques such as the ROCK algorithm.

We chose to examine the CLOPE algorithm because it appears that transactional datasets have much in common with folksonomy data. Both deal with large, sparse datasets featuring high-dimensionality and potentially noisy data. Hence, it seemed appropriate to compare the CLOPE algorithm with approaches normally applied to document clustering.

Unlike K-means and the hierarchical algorithms, the CLOPE algorithm also attempts to optimise the number of clusters produced.

Thus, instead of modifying the algorithm to produce fixed cluster sizes, we varied the *repulsion* variable to produce clustering solutions of different sizes.

8.5 RESULTS

We ran each of the clustering algorithms on the two datasets and measured ICS, and CS for each result. We compare these results to the random baseline for each.

8.5.1 *Intra-cluster Similarity*

Figures 33 and 34 show ICS values for both datasets. In both cases, cluster similarity increases roughly linearly with the number of clusters. All the algorithms show a significant improvement on the random baseline, which follows something like a quadratic pattern. This shows that the folksonomy tags have a strong correlation with the categories for both datasets. If the tags showed no correlation with the categories, then we would see no improvement over random clustering. Instead, we see consistent improvement across the range of cluster numbers.

As expected, the similarity is one when each item is in its own cluster. When the number of clusters is small however, the similarity reaches a minimum value of around 0.28 for the Slashdot dataset and 0.12 for the RawSugar dataset. This is because most of the items in the dataset contain some similarity to at least one other item. Hence, there is a residual ICS within the dataset as a whole (as discussed in Section 8.4.2).

In Figure 33, there is not a great deal of difference between each of the algorithms relative to the random baseline. In Figure 34 however, there is a more distinguishable difference between the four algorithms; the ROCK algorithm clearly shows greater similarity values for cluster numbers greater than ~600, and the CLOPE algorithm shows lower values for cluster numbers less than ~1200.

8.5.2 *Category Scatter*

Figures 35 and 36 show the inter-cluster category scatter measurements for both datasets. Again, all four algorithms show a significant improvement over random assignment. In particular, the ROCK algo-

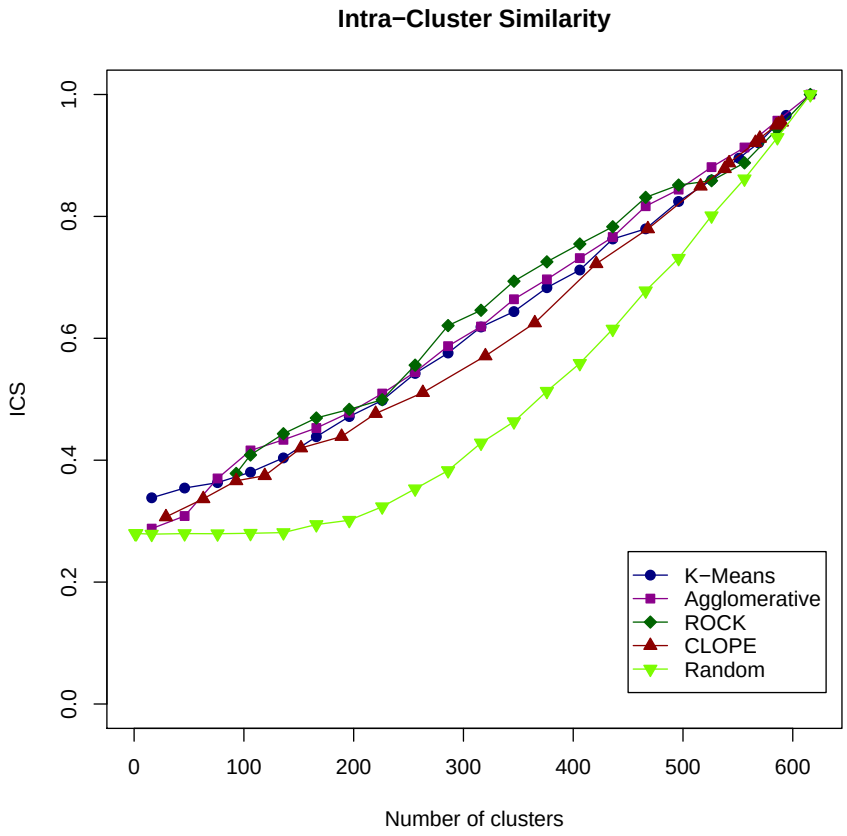


Figure 33: Intra-cluster similarity versus number of clusters for the Slashdot dataset.

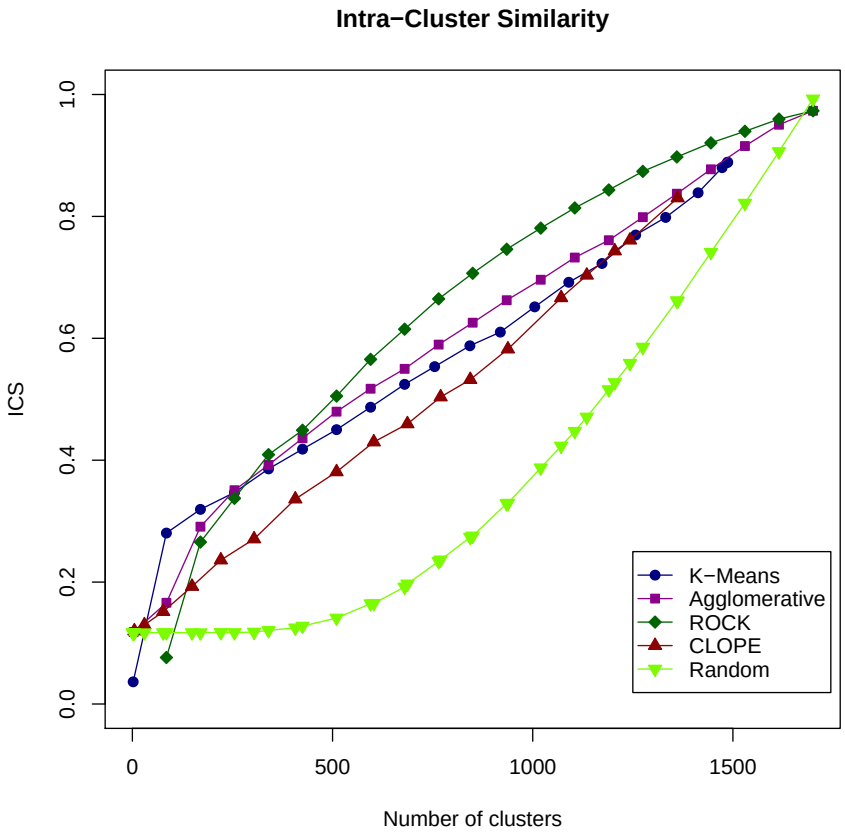


Figure 34: Intra-Cluster Similarity versus number of clusters for the RawSugar dataset

rithm consistently produces less scatter across the range of values tested.

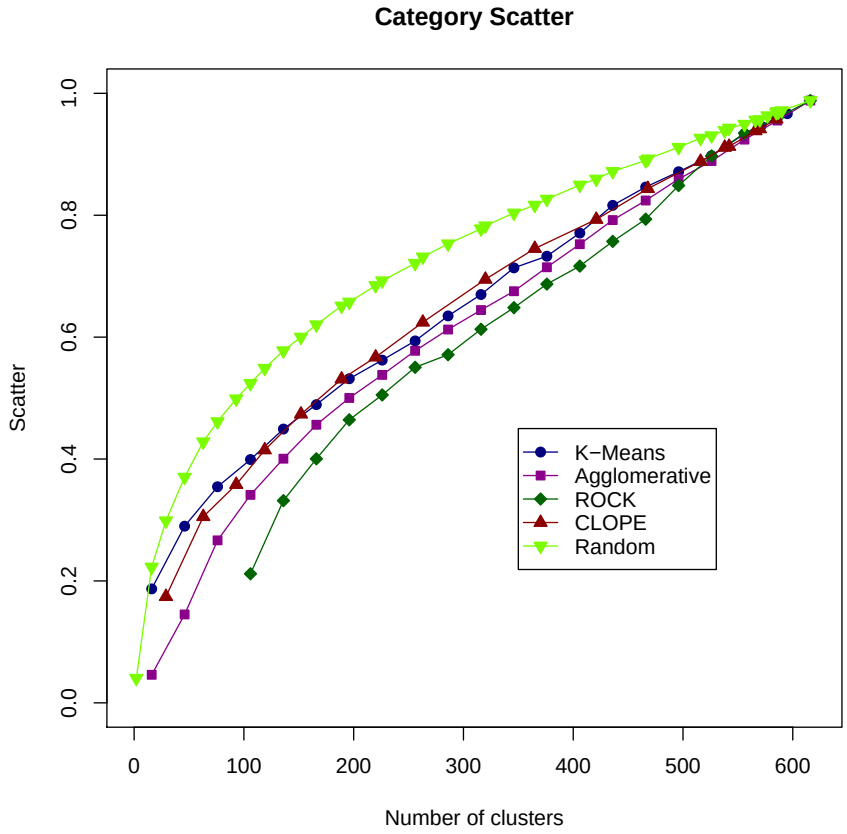


Figure 35: Category scatter for each clustering algorithm. Lower is better.

In figure 36, the difference in scatter is much more pronounced. This is due to the larger number of categories in the RawSugar/D-MOZ dataset. In addition to this, the performance improvement offered by the ROCK algorithm is also more pronounced.

8.5.3 Q *Measure*

Figures 37 and 38 show the raw Q measure scores for each of the clustering algorithms and the random baseline. In both figures, we can see that the ROCK algorithm clearly out-performs all the other clustering algorithms across most cluster numbers. This is especially pronounced in Figure 38. We hypothesise that this is because the ROCK algorithm not only takes into account shared tags, but also common neighbours.

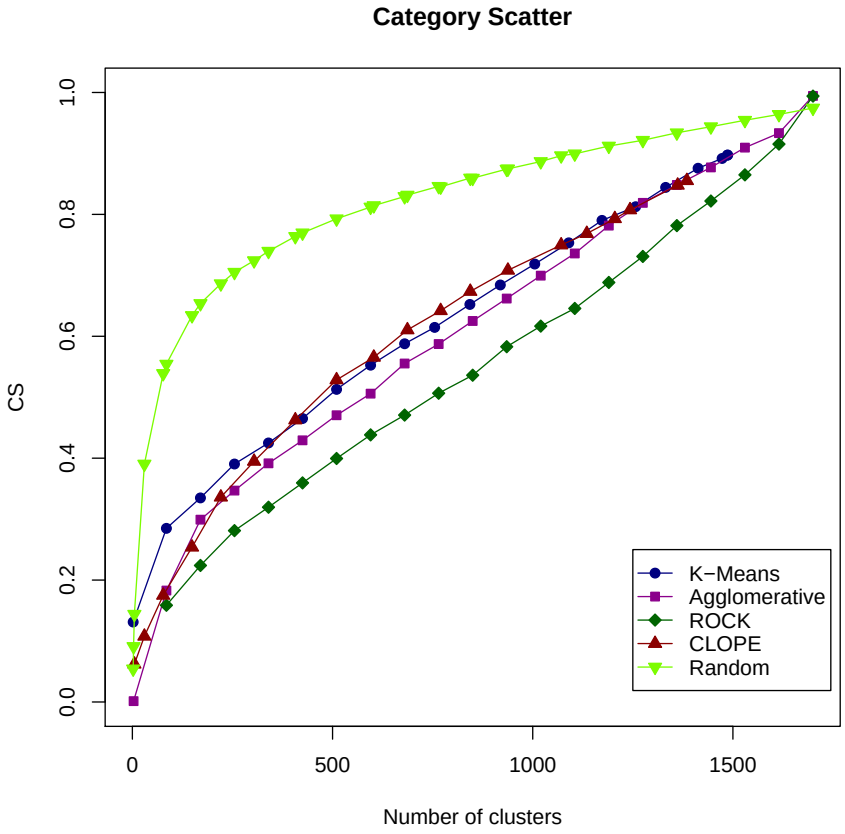


Figure 36: Category Scatter for the RawSugar dataset. Lower is better.

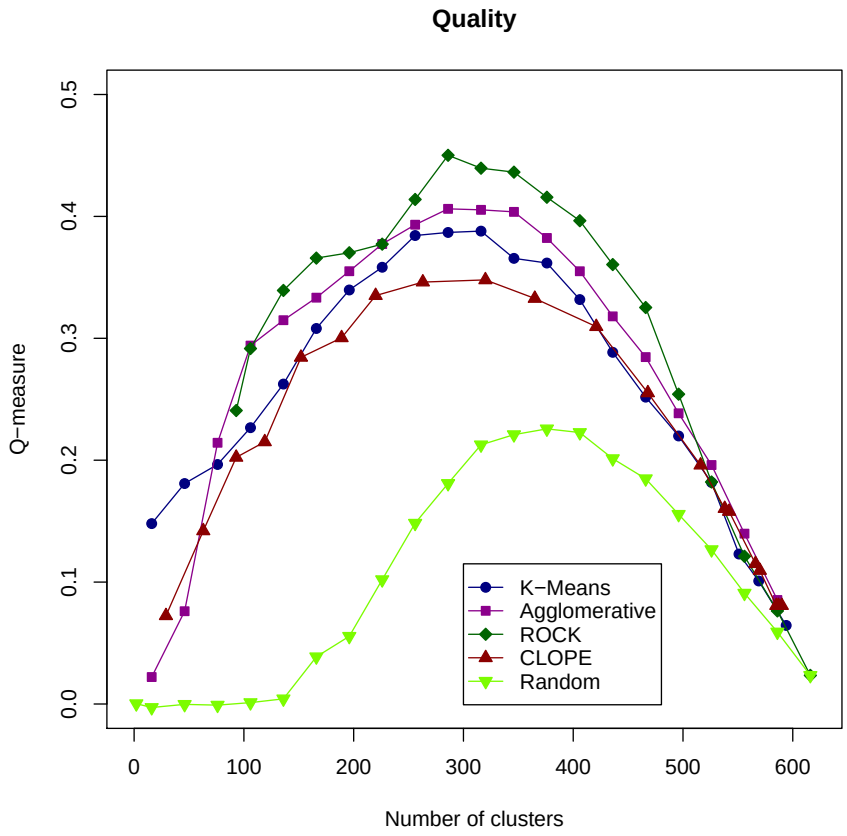


Figure 37: Raw quality measure for each algorithm and the random baseline

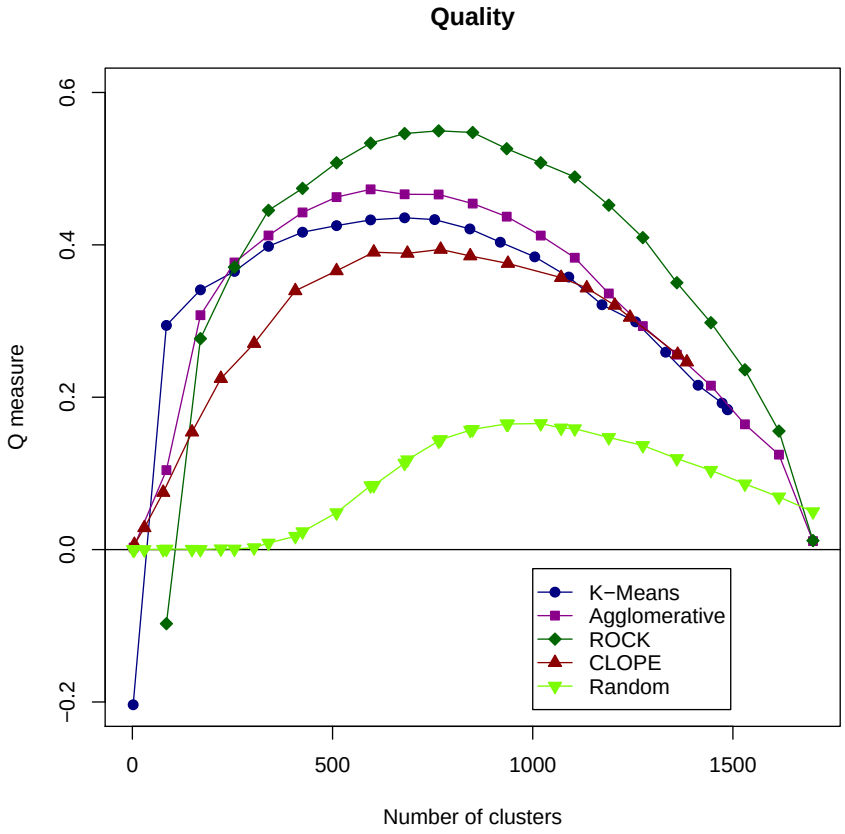


Figure 38: Q-measure for the RawSugar dataset.

The CLOPE algorithm performs worst overall for both category similarity and category scatter. One disadvantage of the CLOPE algorithm is that maximising the height-to-width ratio for each cluster fails to take into account tag scatter. That is, so long as the height-to-width ratio doesn't change, there is no penalty if very similar items are spread across different clusters.

8.5.4 *Clustering Results*

While numerical measurements are a useful tool for examining the effectiveness of a clustering algorithm, they are useless unless the algorithm produces meaningful output. Hence, it is important to examine the actual groupings produced by the four algorithms. Appendix E shows the ten largest clusters for the Slashdot dataset when each clustering algorithm is run to produce the optimal number of clusters. We have listed the frequency of each tag and category in the cluster that occurs more than once, along with the size of the cluster. From this, we can examine how the larger clusters produced by each algorithm compare with the Slashdot categories.

When looking at the clustering results, it is immediately apparent that the largest cluster for each algorithm centres on the tag 'space' and has a high correspondence with the Slashdot category 'Space'. The ROCK algorithm groups these together most strongly, creating a cluster of 105 articles. All of the algorithms feature clusters on related topics, however, and are centred around tags such as 'Mars', 'NASA', 'Moon' and 'telescope'. The reason ROCK performs so strongly may be because it tends to group like items together, even if this results in a particularly large cluster (cf. Appendix E). For this application, such a clustering is appropriate because some categories are much larger than others (e. g. *space* and *NASA*). In some cases however, uneven cluster sizes may be undesirable.

8.6 DISCUSSION

The improvement offered over random assignment shows that clustering based on tags can identify meaningful clusters. If the tags in the folksonomy dataset had no correlation with the categories we measured against, then we would expect that none of the clustering algorithms would provide much improvement over random assignment. Instead, we find that all of the algorithms tested were able to

identify meaningful clusters.

As with most design considerations, there is a trade-off between computational complexity and the quality of results. Our results showed that the ROCK algorithm provided the greatest quality improvement over random assignment; however, it must also be noted that the ROCK algorithm was also the most computationally expensive. Whether or not this is important depends on the application where the clustering algorithm is used. In situations where clustering is performed periodically to create a static dataset, the time taken to complete the clustering may not be an issue. For a dynamic application where clustering is performed in response to a user request however, speed considerations may outweigh those of quality. The algorithms tested here are just a few of the many clustering algorithms reported in the literature. There have been many variations and improvements on the K-means algorithm and agglomerative hierarchical clustering. The number of options to be considered is endless. Our investigations here have provided a starting point for further investigation when considering which algorithm best suits a particular folksonomy-based application.

8.7 CONCLUSION

Many people have begun to apply clustering techniques to folksonomy datasets in a variety of applications. In the literature however, very few authors have given detailed justification for their choice of clustering technique. Hence, we have begun investigating which clustering algorithms are best suited to folksonomy-based applications.

In this paper we have compared the performance of a number of clustering techniques on a poorly conditioned folksonomy dataset. Our results showed that the ROCK algorithm gives the best quality of clusters, but is also the most computationally expensive. Hence, the needs of the particular application must be examined before choosing a particular technique.

We also introduce a new method for comparing clusters with a categorisation scheme where items may belong to more than one category, that is, category membership is not mutually exclusive. Unlike other cluster validity measures, our method does not have a strong bias towards very small or very large cluster sizes.

9

DESIGN OF A FOLKSONOMY-BASED SYSTEM

Chapters 7 and 8 showed that tag clouds and clustering are useful techniques for presenting folksonomy data in a small–medium sized information system. How does one use these techniques however? What does a folksonomy-based system using tag clouds and clustering look like? This chapter documents the design of a folksonomy-based system implemented at the ANU, showing how folksonomies can be used to organise information in a small–medium sized information system.

By documenting the design of a concrete system, this chapter shows how folksonomies can be used practically. This chapter also discusses the potential for folksonomies to support social interaction and knowledge sharing.

9.1 MOTIVATION

The information system described here was designed to assist graduate research students in the Department of Engineering at the ANU to keep track of bibliographic references. A typical Ph.D. student, for instance, will read a great many journal articles, conference papers and books over his/her course of study. When it comes to writing a dissertation, it can be quite difficult to recall a paper read three years ago. To address this issue, I created a system to help graduate research students keep track of papers they have read.

The Department of Engineering at the ANU conducts research in a variety of different disciplines, reflecting the multidisciplinary nature of modern engineering. Research students in the department work in a range of research areas including manufacturing systems, materials research, semiconductor devices and solar cells, telecommunications and signal processing, and control systems. In spite of the diversity of subject material however, ther is a surprising amount of overlap in the methods and techniques used by different disciplines. One of the aims of the system is to help students discover these areas of overlap and build of each others' respective strengths. This kind of discovery is social in nature, hence I called teh system 'SocRef', short for 'Social Reference Manager'.

In a multidisciplinary department like this, people write articles

for many different conferences, journals and other publications. The various publications often have quite strict formatting requirements for both document text and bibliographies. Futhermore, people in the department also use a variety of different systems for inserting citations into their documents. By providing a web-based system, SocRef allows students to organise bibliographic references independently of the format of document they are writing.

People in the department tend to use two main reference management systems: EndNote and BibTeX. EndNote is used primarily with Microsoft Word formatted documents, while BibTeX is used primarily with the LaTeX document formatting system. Different journals and conferences often have specific requirements on the format for acceptable submissions—some requiring Microsoft Word documents, others preferring LaTeX. This can make life difficult if someone has collected their references in BibTeX, but is required to submit a Microsoft Word document. SocRef provides a solution to this issu by providing an independent system which can export references into a format suitable for either management system.

In addition to providing a platform-independent system, SocRef aims to encourage students to share resources with each other and to discover other students researching in similar areas. To facilitate this, SocRef allows users to browse bibliographic references in other people's collections. This allows people to get an idea of others' research interests and potentially discover people working in similar areas.

9.2 RELATED WORK

The idea of using a folksonomy-based system to manage references is not particularly new. A number of web based reference management systems already exist to help people manage bibliographic refernces. Thes include CiteULike*, Bibsonomy† and Connotea‡. Each of these allows people to tag their own bibliographic references, view others' collections and export to BibTeX or RIS. While it may have been possible to simply encourage resarch students to use one of these services, this would not have served the purpose of sharig resources specifically within the Department of Engineering. Systems like CiteULike and Bibsonomy are quite large and it would have been easy for the Department's students to become lost amongst the mileu of

* http://www.citeulike.org
† http://www.bibsonomy.org
‡ http://www.connotea.org/

oher users. Creating a system custom built for the department allows the system features to be specifically targeted towards sharing knowledge within the department.

Another reason for developing SocRef was to demonstrate the use of clustering and tag clouds to support information retrieval in a small–medium sized information system. SocRef is not the first folksonomy-based system in the world to utilise tag clouds and clustering; however, other sytems tend to use tag coulds and clustering differently from SocRef. Notably, the Flickr photo sharing system uses both of these. However, neither tag clouds nor clusters appear to be a primary feature of the Flickr user interface. The tag cloud for Flickr is present on only one page, and appears to be presented as something of a novelty, an interesting visualisation, rather than a navigational tool. Clusters in Flickr are only presented as a secondary link from the search results interface. Again, they appear to be presented as a novelty, rather than as a primary navigation tool. In addition to this, Flickr is a very large information system. It has millions of users and photographs. For the purposes of this thesis we are interested in the application of folksonomies to systems much smaller than Flickr.

Another system to use tag clouds and clustering is the groop.us system developed by Bielenberg and Zacher [14]. In their system, they clustered the resources in each user's collection, then presented these clusters as a set of tag clouds. Each cluster represents the system's guess at what a person's key interests might be. They also use these clusters to build what they call *foci*, which represent a common interest amongst a number of different people using the system. This use of clusters has a different emphasis from the way clusters are used in SocRef. In groop.us, clustering is used to identify shared interests amongst users, while in SocRef clusters are used to assist in disambiguating tags.

Bielenberg and Zacher [14] also use tag clouds in a different way from SocRef. In their prototype system, tag clouds are used to represent clusters and foci, rather than entire collections. They also chose to create circular tag clouds (e. g. Figure 39), with more frequent tags positioned close to the centre of the circle, and less frequent tags at the periphery. This is quite different from the use of tag clouds as described in Chapter 7 and implemented in other systems such as Flickr, Delicious and Technorati. SocRef implements a more conventional use of tag clouds and shows how they might be used to facilitate navigation in a small–medium sized information system.

Figure 39: Example of a circular tag cloud [14, p. 74].

9.3 USING SOCREF

To give an overview of SocRef from a user perspective, we will use a *system-in-use story*. This is a fictional example of a single, specific actor using the system [26]. A system-in-use story gives the description of the system a context and motivation that helps to demonstrate the usefulness of the system. In this case, we shall describe a hypothetical Ph.D. student, Albert.

Albert is in his second year researching materials science, and has already written a three conference articles. He recently started using SocRef after attending a seminar where the author explained its benefits. In this section we shall describe how Albert uses SocRef to browse resources, and to aid in writing a research paper.

9.3.1 *Browsing*

When Albert navigates his browser to the SocRef home page, he is presented with a login form as shown in Figure 40. The login form asks Albert to enter his university ID and password. This username and ID are the same ones Albert uses for other IT systems around the university so he is used to providing this information.

Since Albert has used SocRef before, the system checks his password against a local, encrypted copy of his password and takes him straight to the 'What People are Reading' page. If he had not used SocRef before, then SocRef would have connected to the university

Figure 40: The login page for Socref

LDAP server to confirm his password and download his user information (first name, surname, email address). He would then have been presented with a privacy statement to explain how the information he enters into the system is used.

Normally, the first page Albert sees after logging in is the 'What People are Reading' page (see Figure 41). This page shows the most recent entries into SocRef made by everyone using the system. For example, the first entry in Figure 41 is an entry titled *Studies of diffused phosphoruse emitters: Saturation current, surface recombination velocity, adn quantum efficiency*. Underneath the title, the authors are listed in red, and the first few lines of the abstract give an idea of what the article is about. Beneath the abstract, Albert can see that the article was posted by Thomas, and she gave it the tags *phosphorus emitters* and *silicon*. Albert can also see that Marie has posted quite a few entries recently.

On the right of the screen is a tag cloud allowing Albert to see at a glance the kinds of things people using SocRef are studying. Because Albert is interested in composites, he clicks on the *composite* tag in the tag cloud, bringing up the tag page (see Figure 42).

The tag page shows the most recent entries given a particular tag, in this case *composite*. The display is similar to the 'What People are Reading' page, but a few additional features are displayed on the right. The tag page here shows that the system has found two clusters for the tag *composite*, one relating to *sandwich* structures, *impact*, and *aluminium foam*, and another related to *forming* and *thermoplastics*.

Figure 41: The 'What People are Reading' page.

Figure 42: The tag page.

Below this, SocRef displays a list of people using this tag, including Albert himself.

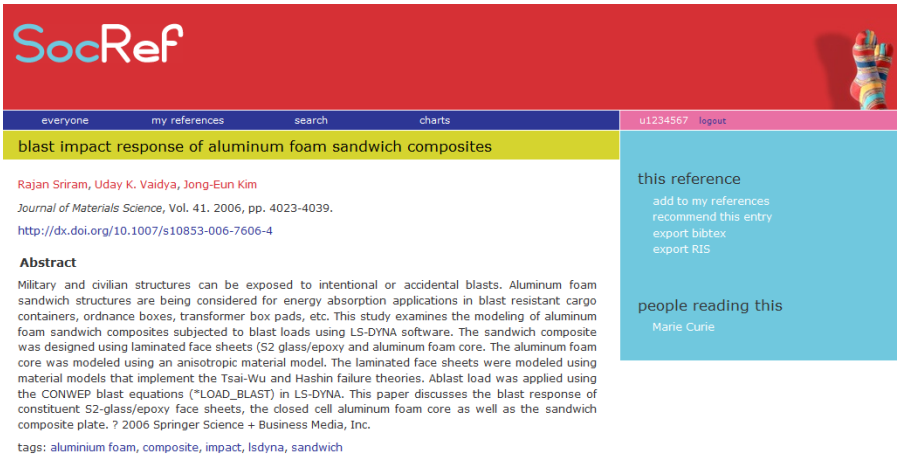


Figure 43: Page showing a summary of information for a single resource.

Albert is interested in sandwich structures, so he could refine his search further using the first cluster.In this case however, the first entry on the page is relevant to him, so he clicks on the title *Blast impact response of aluminum foam sandwich composites*. This brings up the summary page showing the full abstract of the article, and the details of where it came from (see Figure 43). The page also conveniently displays a URL for the paper. The article appears quite relevant to Albert's research, so he opens the link in a new window and prints a copy. He also clicks the 'add to my references' link on the right of the page, adding the article to his collection.

Clicking up the 'add to my refereences' link adds the article to Albert's collection and brings up the page for addign tags and comments (Figure 44). On this page, Albert is asked to enter tags for the article. SocRef suggests some potentially relevant tags, allowing Albert to add them with a single click. Albert adds *sandwich*, *impact*, *composite* and his own tag *energy absorption*, then submits the form. This brings him back to the summary page, now showing him slightly different menu options because the article has been added to his collection. From here he can export the citation information to BibTeX or RIS format for inclusion into a paper, or recommend this entry (via email) to someone else who uses SocRef.

This example shows how a tag cloud can be used to browse a folksonomy-based dataset and also promote sharing of resources. In

Figure 44: User interface for entering tags, showing system-suggested tags

the example, it was quite clear that Albert shared a common interest with Marie in researching composite sandwich structures. The tag cloud allows Albert to see at a glance what kinds of resources are in the collection, and find articles in his area. The clusters allow him to further refine his search to articles that are relevant to him.

9.3.2 *Data Entry and Export*

An important aspect of SocRef is the ability to easily transfer information in and out of the system. Imagine our Ph.D. student Albert is writing a conference paper. He logs into SocRef as before, and is presented with the 'What People are Reading' page. Because Albert is focussed on writing his paper at the moment, he is not interested in what other people are reading, so he clicks on the 'My References' link at the top of the page. This brings up a page showing the most recent entries into Albert's collection (Figure 45).

At the right of this page is a list of links allowing him to upload new resources or export a group of resources. Below this is a personal tag cloud for Albert (which is not particularly large because he hasn't made many entries). And below the tag cloud is a list of other people using tags that Albert uses.

Looking at his list, Albert notices that he hasn't uploaded one of the articles he needs for the paper. He did however, reference the article previously in another conference paper, and still has the BibTeX

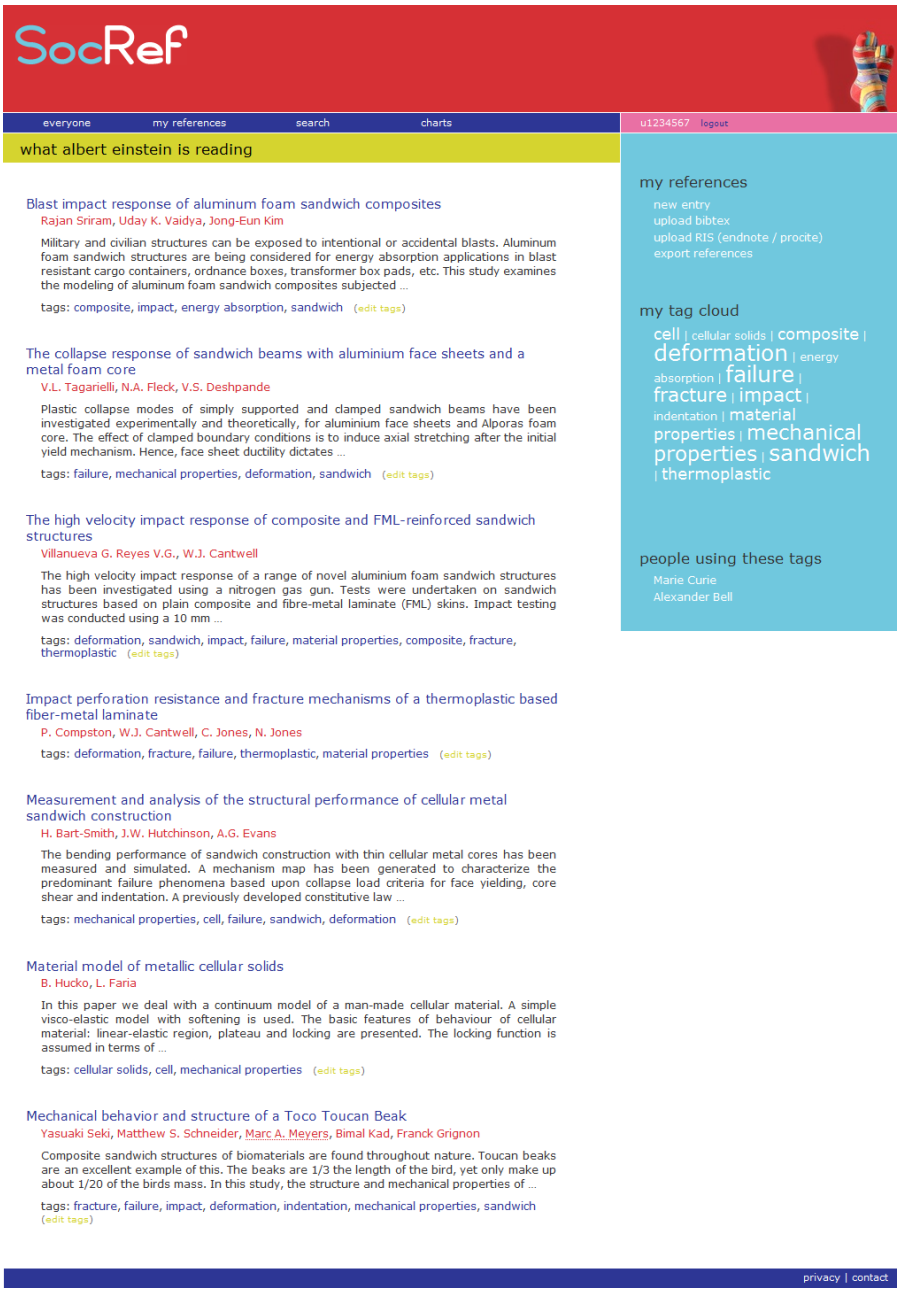


Figure 45: The 'My References' page

file stored on his computer. To upload the article details, he clicks on the 'upload bibtex' link on the right, bringing up the page for entering BibTeX information (Figure 46). He then opens up his BibTeX file, copies and pastes the entry into the text box, then clicks 'Submit'.

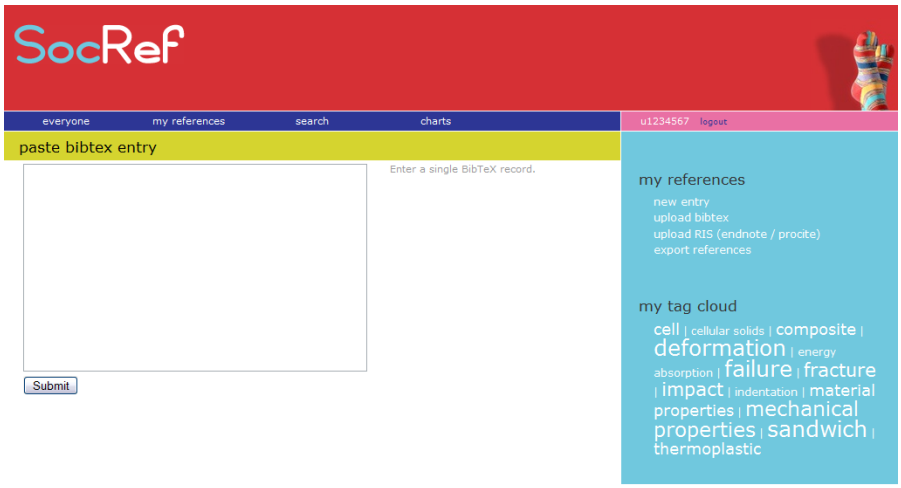


Figure 46: Interface for uploading a single BibTeX entry

The system again brings up the page for entering tags (Figure 44) along with some auto-suggested tags. He clicks on one or two of these suggestions, adds his own tag *sandwich*, then clicks 'Submit' again. This brings Albert to a summary page for the resource, so he can check that the details have been uploaded correctly. Observing that everything seems fine, he then goes back to the 'My References' page.

Now that the article has been uploaded, Albert wishes to begin putting citations into his paper. The papter is required in Microsoft Word format, so he is using EndNote to format his bibliography. From the 'My References' page, Albert clicks the 'export references' link, bringing up a form with a few options (Figure 47). Albert chooses to export all his references with the tag *sandwich* into RIS format. This creates a file which is opened directly by EndNote, allowing him to conveniently add citations.

If Albert were to notice that one of his resources from SocRef had not been exported with the others (i. e. if it did not have the tag *sandwich*), he can simply bring up the summary page for that resource, then click 'export RIS' to create another file that will open in EndNote.

Albert can now use EndNote to insert citations into his document and automatically format the bibliography. Using the export functions of SocRef, Albert was quickly able to transfer relevant citation data from SocRef to EndNote. If he was writing a document in LaTeX,

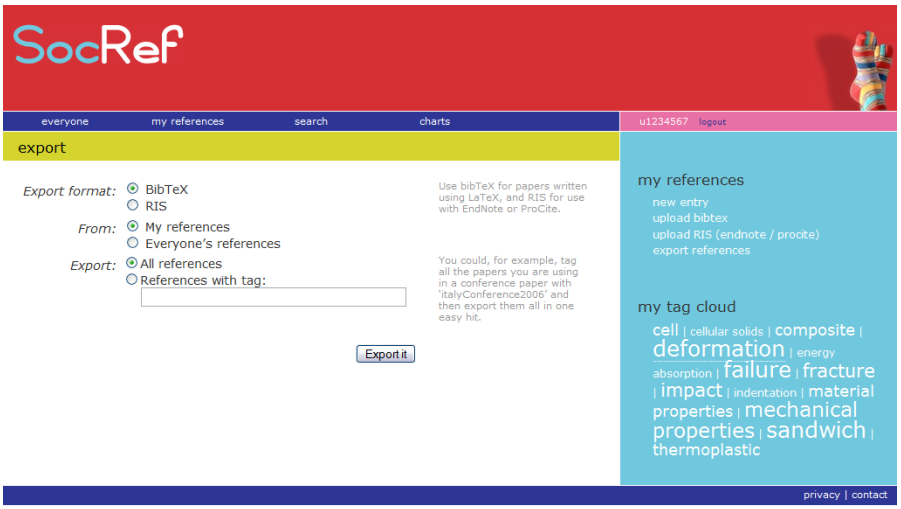


Figure 47: Options for exporting a group of resources

the procedure for exporting to BibTeX would have been much the same.

9.4 SYSTEM DESIGN

Having described how people can browse SocRef and import/export reference information, we now turn to look at the design of SocRef in more deatil. In particular, we examine the system architecture, data entry processes, and data presentation aspects of the system.

9.4.1 *System Architecture*

In the SocRef system, like any folksonomy, there are three primary entities represented: people, resources and tags. These three entities form a tripartite network [74]. A person will have a number of resources in their collection which are labelled by tags. This relationship is visualised in Figure 48.

Each person using SocRef has their own collection of resources. SocRef records each person's university ID (as a unique identifier), their first name, surname, and email address. The email address is recorded so that people using SocRef have the ability to recommend resources to each other via email.

People using SocRef are identified by their real names, rather than a pseudonymous username as with systems such as Delicious or Flickr. This is because one of the aims of SocRef was to promote interaction between people sharing common interests. If I discover someone who

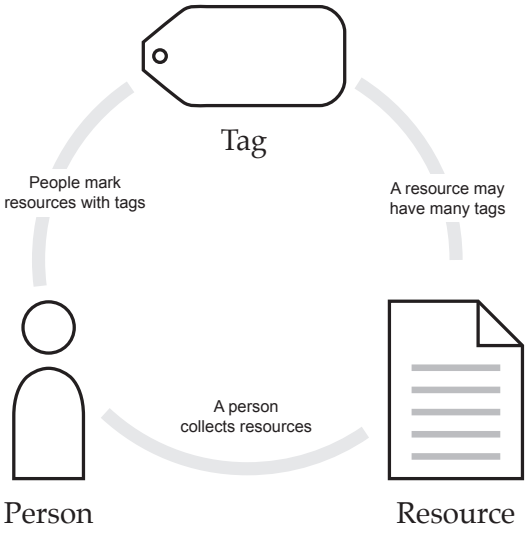


Figure 48: People, Tags and Resources form a tripartite network

shares a common interest with me through SocRef, then, knowing their real name, I can talk to them personally about their research.

A resource in SocRef represents anything that might be cited in an academic paper or book. This might include journal articles, conference papers, books or websites. Different reference managers have different sets of allowable resource types. Since SocRef aims to be independent of different reference managers, it defines its own set of allowable resource types[§] that is something of a compromise between BibTeX and RIS. These types are listed in Table 38, along with their corresponding types in BibTeX and RIS formats.

To support such a broad range of resource types, SocRef allows people to enter many different kinds of information. The field types and descriptions are listed in Table 39. Each resource requires at least a citation type and a title; all other fields are optional.

Tags in SocRef can be any string of characters that does not contain an illegal character. Illegal characters include `'`, `"`, `+`, `|`, `(`, and `)`. This means that tags may include white space (unlike many other tagging services), removing the need to form a convention on how to format multi-word tags.

9.4.2 *Entering Information in SocRef*

Entering bibliographic information can be a tedious process. One of the aims of SocRef was to make this process as quick and simple as

§ The list given here is loosely based on a list created by Dana Jacobsen [63]

Table 38: Different types of resource

TYPE	DESCRIPTION	BIBTEX	RIS
article	An article from a periodical (journal, magazine, newspaper).	article	JOUR
book	A standalone book.	book	BOOK
booklet	A booklet or pamphlet.	booklet	PAMP
collection	A collection of works, not proceedings.	book	BOOK
inbook	A part of a book or collection given by a chapter number or range of pages.	inbook	CHAP
incollection	A part of a book or collection with its own title.	incollection	CHAP
proceedings	The proceedings of an event such as a conference.	proceedings	BOOK
inproceedings	An article from the proceedings of an event.	inproceedings	CONF
manual	Technical documentation.	manual	RPRT
thesis	Masters or Ph.D. dissertation.	mastersthesis, phdthesis	THES
report	A technical report.	techreport	RPRT
unpublished	A letter, treatise, manuscript, or other unpublished work	unpublished	UNPB
avmaterial	Musical scores, Film, Art, etc.	misc	GEN
misc	Anything at all.	misc	GEN

Table 39: Field types for a resource in SocRef

FIELD	DESCRIPTION
`ID`	An ID for the resource
`title`	Title of a work. The title of the work being referenced, not the collection or book it is found in (use `superTitle`).
`superTitle`	Title of the object that contains the record.
`journal`	If an article, the title of the journal, magazine, or newspaper that the article appears in.
`volume`	The volume, either of a journal, or of a book.
`number`	The number of a journal. `ReportNumber` is used for reports.
`series`	If a book, journal, proceedings, etc. is in a series, this is the name of the series.
`edition`	The edition or revision (e. g. 'Second').
`chapter`	If `inbook` or `incollection`, a number for the chapter. This is not the chapter title (use `title`).
`pages`	The pages on which the resource is found.
`publisher`	The name of the publisher. Should not contain the address.
`pubAddress`	The address for a publisher.
`month`	The month that the resource was published (or for unpublished works, when it was written).
`year`	The year the resource was published, or written for unpublished works.
`day`	The day the resource was published.
`reportType`	The type of report or dissertation (e. g. 'Technical', 'Masters', 'PhD').
`reportNumber`	The number of a report, if it has one.
`abstract`	An abstract of a resource.
`note`	Any other information that should appear in a bibliography.
`organisation`	The organization or institution associated with the work. For technical reports, this is the organisation sponsoring the work. A thesis does not use this field—it uses the `school` field instead.
`school`	The school where a thesis was granted.
`ISBN`	The ISBN of a work. This is formatted in the string 'c-ccccc-ccc-c' where the 'c' values are either digits or 'X'. A record can have multiple ISBN numbers, comma seperated.
`ISSN`	The ISSN of a work. This is usually associated with a journal rather than a particular article. This is formatted in the string 'cccc-cccc' where the 'c' values are either digits or 'X'. Multiple ISSN numbers are comma seperated.
`copyright`	A copyright message.
`howPublished`	How something is published if it is unusual.
`url`	A pointer to the actual information this record references.
`language`	The language the work is in.
`citeType`	The type of resource (see Table 38)

possible, otherwise there would be little personal advantage for students wishing to use it. In the early stages of development, surveys and interviews with research students in the department revealed that they normally located references through online databases such as Engineering Village[¶], Science Direct[‖] and Proquest 5000[**]. Most of these databases provide a facility for exporting citation information in either BibTeX or RIS format. To simplify the data entry process, SocRef allows users to copy and paste BibTeX or RIS into the browser and upload this information in a single click. This means that users are able to quickly enter information from databases, or their own personal data files with relative ease.

In some cases BibTeX or RIS information is not available however, so SocRef also provides an interface where reference information can be entered manually (see Figure 49).

In summary, SocRef provides three different methods for entering resource information: 1) pasting BibTeX; 2) pasting RIS; and 3) manual entry. After resource information has been entered into the system using one of these methods, SocRef analyses the reference information in order to suggest appropriate tags. SocRef then prompts the user to enter tags for their resource.

Both RIS and BibTeX formats must be parsed in order to convert them to the canonical format described in Section 9.4.1. In the case of RIS, this is performed using a series of regular expressions. The BibTeX format is slightly more complicated however, so the SocRef system runs BibTeX itself on the input data, converting it into a form suitable for entry into SocRef.

None of the three data entry pages allows multiple resources to be uploaded at once. This was a deliberate design decision as uploading entries one at a time allows students to consider the tags they enter more carefully than if they simply uploaded 100 or so resources at once. Unfortunately however, this does make the data entry process more tedious.

After a resource is entered, SocRef analyses the information in order to suggest tags. It does this using two methods. The first simply scans through the title and abstract fields of the resource looking for any words that match tags already in the system. The second mechanism takes the author, title and abstract fields and calculates the similarity of the resource to all the other resources in the database.

¶ http://www.engineeringvillage2.org/
‖ http://www.sciencedirect.com/
** http://www.proquest.co.uk/

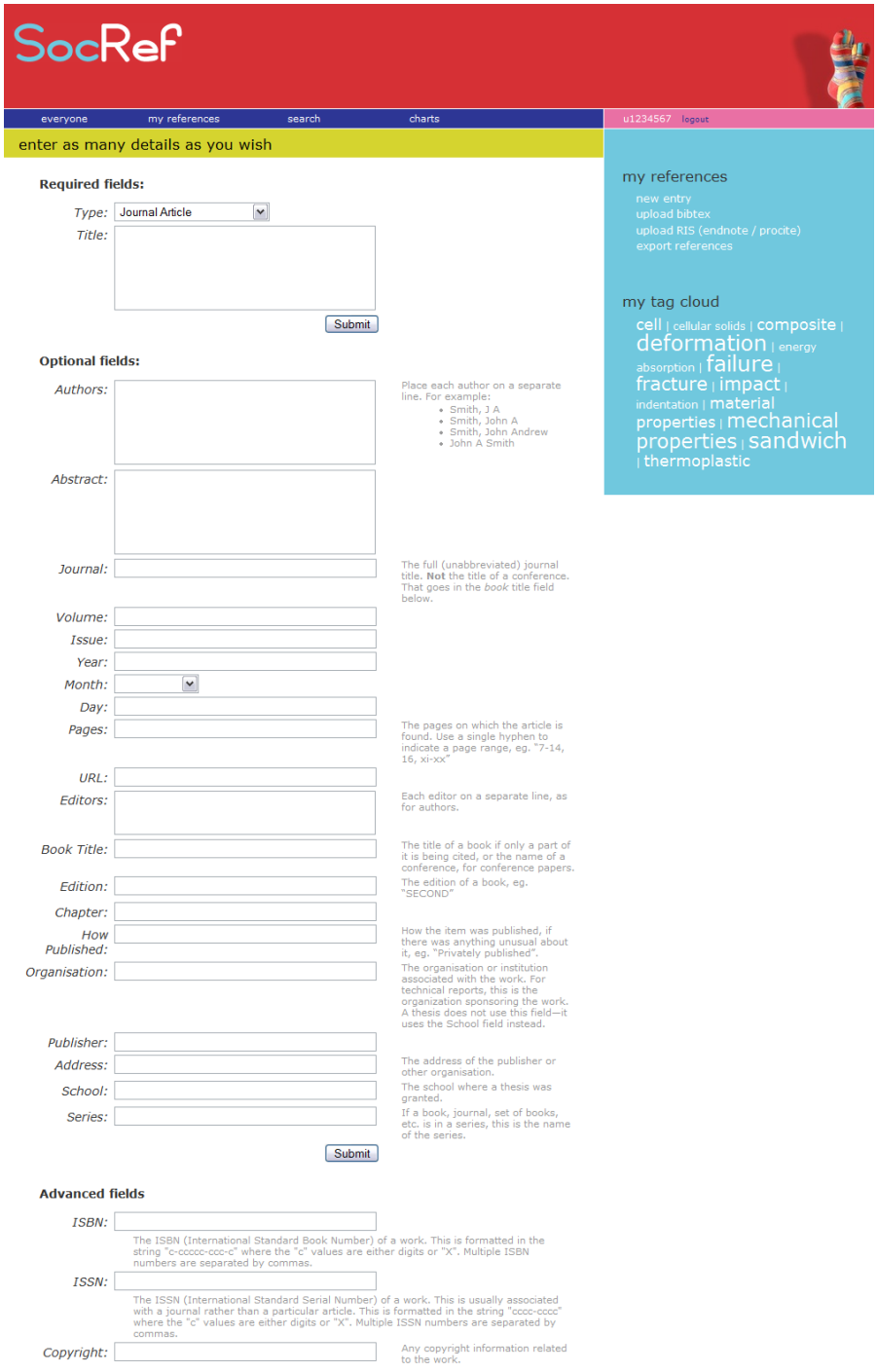


Figure 49: Manual resource entry interface

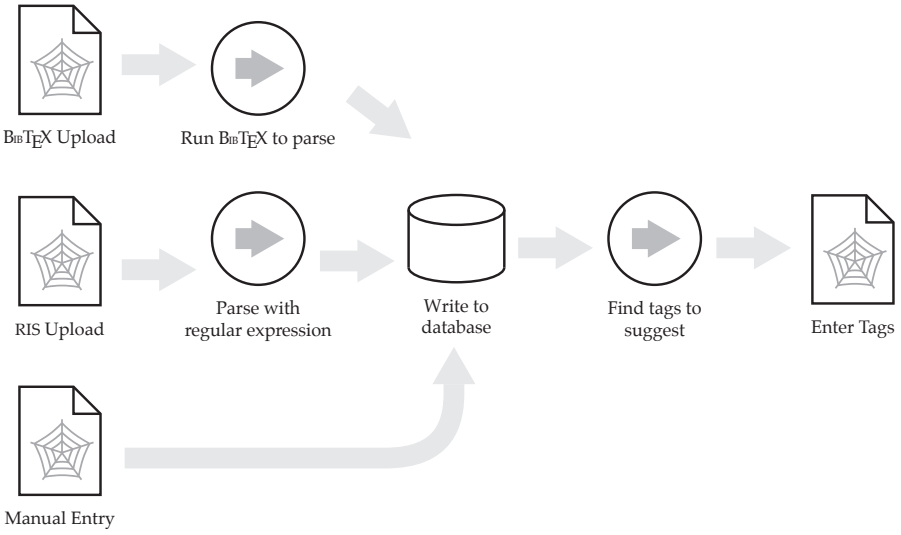


Figure 50: The process of entering a resource into SocRef

If any of the other resources are above a given similarity threshold, the system suggests the tags associated with those resources. The results of these two methods are combined to form a single list of tag suggestions.

Similarity between documents is calculated based on TF-IDF weighting and the SIM4 similarity measure as described by Segal and Kephart [118].

After determining appropriate tags to suggest, the system presents these on the page for entering tags (Figure 44). The same page also allows people to add any personal comments or annotations for the resource.

After tags are submitted, the system presents a summary of the information submitted, as shown in Figure 43. This allows the user to quickly check that the information has been uploaded correctly.

9.4.3 *Finding Information in SocRef*

There are three different approaches that SocRef facilitates to allow discovery of resources: General Browsing, Tag Browsing, and Search. Each of these approaches allows different levels of specificity in finding information.

General Browsing

After someone has logged in to SocRef, they are presented with a page entitled 'What People are Reading'. This page shows the ten

most recent entries made into SocRef. For each entry it lists the title, authors, the abstract, who posted the item, and the item's tags. The title is hyperlinked so that clicking on the title brings up the summary page for that item (Figure 43). This allows anyone to quickly see what people have entered into SocRef recently.

The authors, tags, and who posted the item are also hyperlinked on the 'What People are Reading' page. Clicking on any one of these links brings up a page listing resources by authors, tags or people respectively. Hence, if an article appears to be of interest, one can examine any other resources written by the same author, any resources using one of the same tags, or the resources in the owner's collection. Each of these pages uses the same kind of presentation structure. This ability to move between tags, authors, and people is called *pivot browsing* [96, 97].

Each of these pages also shows a tag cloud, described in more detail below.

Similarly to the 'What People are Reading' page, the 'My References' page shows the a list of the most recent resources posted, however it limits these to entries made by the current user. The tag cloud on this page is also restricted to tags entered by the current user.

Tag Browsing

The ability to click on a tag and display a list of resources with that tag has already been mentioned above. In addition to this, there are two extra features that SocRef provides in relation to tags. These are tag clouds and clusters.

TAG CLOUDS The tag cloud is displayed on the right hand side of most pages in SocRef. The top 50 most popular tags are listed in alphabetical order, separated by vertical bars (|). The size of each tag is proportional to its popularity, calculated using the same equation as Chapter 7:

$$\text{TagSize} = 1 + C\frac{f_i - f_{min} + 1}{f_{max} - f_{min} - 1}$$

Where f_i is the frequency of tag i, f_{min} is the minimum frequency in the top 50 tags, f_{max} is the maximum frequency in the top 50 tags, and C is a scaling constant that determines the maximum text size. This gives a tag size relative to the default font size set by the browser.

Clicking on any tag in the tag cloud brings up a page showing the

most recent resources entered with that tag, and any clusters associated with the tag.

CLUSTERS Clusters are displayed in the top right corner of the tag page. Each cluster is represented by the most frequent tags that occur in the cluster. Clicking on a cluster link brings up a list of resources in that cluster.

Clusters are generated using the ROCK algorithm described in Chapter 8. For each of the top 50 most popular tags, SocRef selects all the resources associated with that tag, and attempts to find an optimal clustering. Once the optimal number of clusters has been determined, the system clusters resources to this number, and removes any clusters with less than three items. If more than one cluster remains after small clusters have been removed, then the cluster information is written to the database.

The optimal clustering is determined using a metric similar to the Q-measure described in Chapter 8. Instead of calculating category ICS and CS using an external category system however, it calculates ICS and CS using the tags themselves.

Currently, the clustering algorithm is run every time a person enters tags for a new resource. Since the system currently contains only around 500 records, the clustering does not require too much time to run. Once the system gains more entries, the clustering algorithm may have to be applied more sparingly.

Searching

In addition to pivot browsing, tag clouds and clusters, SocRef also provides a search interface. This allows a user to search for resources with particular words in the title, abstract, or authors. Users are also able to limit the search to only their own references. Search results are presented in a fashion similar to the 'What People are Reading' page.

9.4.4 *Sharing in SocRef*

SocRef also includes a number of features to encourage people to share resources and discover shared interests.

The first of these features is a list of people who use the same tag on each tag page. This allows a user to quickly see a list of other people who are potentially interested in a similar topic. Clicking on any

name in the list brings up a user page listing that person's resources and tags.

Similarly, on each person's page a list of people who use the same tags is given underneath the tag cloud. This provides a list of people who may potentially share similar interests.

Another feature is email recommendation. When viewing the summary for any resource, a link on the right-hand side of the page allows a user to send an email recommending the resource to any other SocRef user. Each summary page for a resource also includes a list of any other people who have the same resource in their collection.

9.5 IMPLEMENTATION

I chose to implement the system using open source development tools that would be compatible with the departments existing IT infrastructure. Specifically, I used a MySQL for the database back-end and PHP scripts to create the HTML front end. Both PHP and MySQL are well established development tools that are used frequently in large scale web applications (including folksonomy-based web applications).

One of the advantages of MySQL is that it provides built in search algorithms for full-text search (described previously in Chapter 7). Unfortunately however, using the full-text search features of MySQL means that foreign key constraints must be disabled. This adds to the complexity of some queries in the system, since foreign keys must be checked manually. These issues are well known however, and do not cause major issues.

The system also makes good use of Javascript code and AJAX techniques making use of the Xajax[††] library. In particular, this allows people to edit their tags on the 'My References' page without having to refresh the page. This makes editing more convenient and improves the useability of the system.

9.6 DISCUSSION

9.6.1 *When is a Folksonomy Useful?*

Describing the design of a system like SocRef demonstrates how folksonomies, tag clouds and clustering might be used to organise information in a small–medium sized system. While this is useful, it

†† http://www.xajaxproject.org/

does not tell us *where* and *when* a folksonomy might be useful. Folksonomies are by no means a solution to every categorisation problem. In what situations then, will a folksonomy be useful?

The first requirement for a folksonomy to be useful is that there is no expert administrator available to organise the information. In most cases, if an individual (or team of individuals) is able to spend time regularly maintaining a category scheme, then a folksonomy is not really necessary. An information architect or librarian can keep the system well organised while avoiding the drawbacks of folksonomies such as homonymy, synonymy and tag noise. In some cases, an information architect may use a folksonomy to supplement their research in developing a category scheme, but it will not usually be the primary method of organsiation.

The second requirement for a folksonomy is that people have some reason to browse other people's collections. In the case of SocRef, browsing another person's collection gives an idea of the kind of research they are working on, and could potentially lead to the discovery of useful resources. In the case of an email program, on the other hand, it is important that people *not* be able to browse others' collections. Unless there is good reason to explore the collections of others, many of the benefits of a folksonomy are lost.

Finally, for a folksonomy to work well, people must have some kind of ownership over the entries they put into the system. For example, if Helen is simply entering data on behalf of others because it is her job, then she has little personal association with the information. There would be little point in allowing others to browse her 'collection' because it has very little to do with Helen. One of the main advantages of folksonomies is that people tend to only enter information that has proved useful to them. This acts as a kind of filtering mechanism, keeping out low quality data.

Having looked at this general issue regaring folksonomies, we now return to discuss SocRef. The following sections describe issues that have arisen from the preliminary implementation of the system. In particular we focus on areas where SocRef could be improved.

9.6.2 *Bulk Uploading*

Feedback from people currently using the system suggests that a number of users would like to be able to upload multiple BibTeX or RIS entries from a single file at once. This would make the data entry process even faster and easier than it is currently. As mentioned pre-

viously, SocRef deliberately does not provide this facility in order to encourage people to think carefully about what tags they use to describe the resource. Unfortunately however, this design decision may have discouraged some people from using the system, as the thought of manually copying and pasting hundreds of entries from their existing database does not appeal.

Initially, the author assumed that people would upload resource information into the system as each article was read. This would lead to one or two entries being made into the system by each user every few days. In reality however, most research students seem to collect bibliographic information together only when necessary; that is, when they must write a paper or thesis. The tendency is to then enter a lot of citation information in a short period of time. This behaviour and feedback from users suggests that a bulk uploading option would make data entry easier for most users of SocRef.

9.6.3 *Clustering Improvements*

The clusters produced by SocRef form a useful method of narrowing a search from the tag page. In most cases, the ROCK algorithm produces nicely separated clusters that fall neatly into semantic categories. Where two tags occur together frequently however, the clusters produced by the ROCK algorithm are not always semantically useful. In the current design, the clustering algorithm mistakenly attempts to keep all these tags together in one cluster, when really they should be separated.

A potential solution to this problem would be to introduce some form of weighting to the tags like the TF-IDF metric. This would make the clustering algorithm less eager to keep all occurrences of a tag together when it occurs frequently with the tag being clustered.

Another potential improvement would be to incorporate personal tag clusters, similar to those implemented by Bielenberg and Zacher [14] in groop.us. This would allow a user to view groups of semantically related resources in their own library and make it possible to find examine common interests between people.

9.7 CONCLUSION

In this chapter we have examined the implementation of SocRef, a folksonomy-based system that allows research students to manage

and share bibliographic resources. In particular this chapter has shown how tag clouds and clusters can be incorporated into a small–medium sized folksonomy-based information system. Tag clouds provide a quick visual summary of everyone's collection and individuals collections and allow for easy pivot browsing. Clustering meanwhile, helps to separate tags into semantically related clusters.

10

SUMMARY, CONCLUSION AND FUTURE WORK

10.1 SUMMARY

Organising information is difficult. Yet information is pushed at us constantly through televisions, radio, billboards, email, reports, meetings, web sites and phone calls. Some of this information is important, useful and relevant; much of it is not. Faced with this flood of information, the challenge is not so much to *find* relevant information, but rather to *ignore* the vast amounts of irrelevant information. To find a needle in a haystack, the trick is to ignore the hay.

In the human mind, categorisation is the mechanism that allows us to focus on what is important. For example, if I categorise an animal as a *tiger*, I can focus on appropriate actions such as *running away* or *hiding*, without paying great attention to the colour of leaves nearby or the amount of dirt under its claws. In information systems, categorisation plays a similar role in hiding what is unimportant. Grouping items together allows me to temporarily ignore everything that is not a member of the group.

Categorisation in the brain appears to happen effortlessly and continuously; it is one of our most basic cognitive functions. In the realm of information systems however, our categorisations seem less than ideal. Cluttered computer desktops and disorganised intranets bear testimony to our difficulties with categorisation. Amongst the clutter it becomes difficult to separate the relevant from the irrelevant. Things may be grouped into categories, but the categories don't fit what we want them for.

This thesis investigates categorisation problems that occur in information systems. The investigation began through a case study of categorisation in a manufacturing environment. The study examined SIMPRESS, a KM system designed to help staff solve problems and prevent future manufacturing issues. The study found many symptoms of categorisation problems, leading to two simple questions: 1) Why do these problems occur? and 2) What can be done about them?

The results from the SIMPRESS study suggested two main causes for categorisation problems: cognitive mismatch and contextual factors.

Cognitive mismatch occurs where the structures in individuals' minds do not match well with the categorisation structures available in the

information system. A survey of the cognitive science literature revealed three key reasons why this might occur:

1. *Misconceptions of categorisation.* One common misconception is confusing categorisation with classification. In a classification system all classes are mutually exclusive; an item belongs in only one 'correct' position in the classification scheme. Categorisation however, is much broader than this. Categories may overlap and items may have varying degrees of membership in a category. Confusing classification with categorisation can cause problems when an item *sort of* belongs to a class, but also *sort of* belongs to another class. Categorisation allows for this kind of ambiguity, whereas classification does not.

2. *Subjectivity and category dynamics.* Categorisation in the mind is highly subjective and dynamic. Categories are created to suit pragmatic purposes specific to the individual's context. That is, categories are created in my mind to suit the kinds of things I do. For example, where most people see a parrot, an ornithologist may see a *rainbowus lorikeetus*. Category structures in the mind are also continually changing as new information becomes available. The mind updates categories unconsciously and effortlessly. Most information systems however, are not as well equipped to deal with category shifts and changes.

3. *The relationship between language and categories.* The link between vocabulary and categories is not straightforward. People choose words to communicate a particular message to a particular audience. Because of this, words cannot be directly correlated with categories, since different words may be chosen in different circumstances. Hence a dog may be an *animal*, a *boxer*, a *pet*, *the neighbour's dog*, or *a stray* depending on what one wishes to communicate.

The three reasons above make the design of category schemes quite difficult because information systems usually have many different people using them. Categories in the mind can be optimised to suit the needs and knowledge of just one individual. In an information system, the category structure must accommodate the different cognitive structures of many individuals. Furthermore, not only do people vary greatly in how they understand concepts, but also in how they describe them.

This potential for variability was shown clearly by the interface study described in Chapter 4. The interface study tested the hypothesis that allowing a graded category structure (i. e. categorisation instead of classification) would increase categorisation accuracy. The results showed that the graded interface did increase accuracy slightly, but overall, participants were not particularly good at predicting the categorisations of their peers.

Cognitive mismatch is enough to make the design of almost any category scheme difficult. However, cognitive mismatch does not account for all of the problems encountered with the SIMPRESS case study. The results from interviews suggested that other factors arising from the system context had a large impact on categorisation. Furthermore, a review of the literature revealed that these issues were not unique to SIMPRESS. These issues include:

CONFLICTING REQUIREMENTS. The different people involved with an information system may have quite diverse purposes and goals. In many cases the people who enter data are not the same people who use the data. The goal of the person entering the data may be to get a tedious task out of the way as quickly as possible, while the goal of the person using the data may be to carry out a detailed statistical analysis. Where goals and purposes differ, there is potential for conflicting priorities.

POLITICAL & SOCIAL CONSEQUENCES. Because we inhabit a social and political world, categorisation has political and social consequences. How people categorise expenditures has consequences for taxation. Building zone categorisations have consequences for residents. When the consequences of categorisation become personally significant, achieving a desired outcome will take priority over accuracy.

PERCEPTIONS OF THE SYSTEM. How people perceive the information system will affect how they interact with it. If data entry is seen as unimportant, then providing accurate categorisations will have a low priority when people are busy and pressed for time.

ENVIRONMENTAL DYNAMICS. Problems occur when the historical categories of an information system are no longer appropriate to the current situation. In such cases people are forced to improvised using localised ad hoc standards and work-arounds. These localised standards will usually suit the needs of the local situa-

tion, but can lead to inconsistency where there are many different groups of people interacting with the system.

PREDICTION. Unless all the items to be categorised are known beforehand, creating a category system always involves predicting the future. When categorisation schemes are created for information systems, usually only a relative few items are available. The designer must predict what kinds of items will be categorised in future and design the category scheme accordingly.

THE TEDIUM OF DATA ENTRY. Data entry is often boring and tiresome. Bored, tired workers tend to make more mistakes and cut corners were they can. Thus, the quality of categorisation suffers.

This review of categorisation problems in the literature shows that the categorisation problem is by no means new or unknown; nor is this thesis the first ever attempt to address these issues. The field of LIS in particular has a long tradition of attempting to engage with categorisation problems. Coming from a librarian's perspective however, writers in this area tend to make a number of assumptions about how categorisation problems will be solved:

1. A dedicated librarian or administrator is available to maintain and modify the categorisation structure;
2. The items to be categorised are primarily textual documents; and
3. The number of items to be categorised is very large.

As illustrated by the SIMPRESS study however, these assumptions do not always hold true.

Taking cognitive factors and contextual factors together gives a holistic view of the categorisation problem. Recognising that the assumptions made in the LIS literature do not always hold true, the aim of this thesis was defined in Chapter 5. That is, the aim of this thesis is to find categorisation methods that:

A. reflect user vocabulary;
B. are able to adapt to change;
C. do not require expert human intervention;
D. are suitable for multimedia data; and
E. work effectively with small numbers of records.

Having defined the problem it is appropriate to see if it has already been addressed elsewhere. A survey of the literature revealed that none of the reported approaches adequately deals with the problem as posed here. However, folksonomies appear to have the most potential for addressing the issues.

Folksonomies are most suited to this particular problem because:

A. They directly reflect user vocabulary;

B. They are able to adapt to changes in vocabulary;

C. They do not require a dedicated administrator; and

D. They are suitable for multimedia data.

The drawbacks with folksonomies are:

1. Folksonomies are not well researched and it is not well known how to use folksonomy data to the best advantage.

2. Folksonomies appear to rely on large numbers of users to obtain consensus, so it is unclear whether they can work in a small information system.

3. Folksonomies suffer from synonymy and homonymy.

The first two of these drawbacks were addressed by evaluating the use of folksonomy tag clouds as a social navigation tool. The results showed that tag clouds are a valuable tool but need to be supplemented with more specific methods of information seeking. As part of the experimental setup, the tag cloud was tested on a small, poorly conditioned dataset. This showed that tag clouds provide value even with a small information system, as specified in the problem definition. Hence, tag clouds are a useful contribution to addressing the problem laid out in this thesis.

Drawbacks (2) and (3) were addressed by applying clustering algorithms to the folksonomy data. The results showed that the folksonomy captures rich semantic data that correlates strongly with human-generated categorisation. The clustering algorithms tested were able to provide useful groupings of articles that could be used to enhance navigation of a folksonomy dataset.

The results of the clustering comparison showed that the ROCK algorithm performs best, however it is computationally expensive of the algorithms. This complexity makes the algorithm impractical for use

in most systems. With the smaller dataset, the hierarchical agglomerative algorithm performed almost as well as ROCK at considerably less computational cost. For the larger dataset however, the difference in quality between algorithms was much more pronounced. Given that this thesis focusses on small–medium sized information systems, the hierarchical agglomerative algorithm may be quite suitable to the task. More research in this area would be useful to determine if better clustering performance could be achieved without the computational cost of the ROCK algorithm.

The studies of tag clouds and clustering algorithms on small datasets shows that folksonomies can indeed be adapted to suit small–medium sized information systems. Folksonomies have already been shown to reflect user vocabulary and respond to change without the need for expert intervention. They are also well suited to multimedia data, as demonstrated by the popularity of services like Flickr. Thus, folksonomies show themselves to be a good solution to the categorisation as outlined in this thesis.

10.2 CONCLUSION

This thesis began as an exploratory study, investigating the causes and mechanisms of categorisation problems in information systems. Looking at both cognitive functions in the brain and the context of information systems showed that categorisation is far from simple. Individuals vary so greatly as to make the design of a perfect categorisation scheme impossible. At the same time however, category structures in the mind are not arbitrary or random, and there are many commonalities between people. Hence a good categorisation scheme will find a balance between accommodating individual differences and encouraging conformity.

The investigation process revealed a gap in the literature concerning assumptions about how to solve the problem. Most proposed solutions assume that an expert administrator is available to maintain category structures, that the items to be categorised will be primarily textual, and that the dataset will be very large. The reality is however, that these assumptions do not always hold true. This thesis proposes that by using tag clouds and clustering techniques, folksonomies can be adapted to suit smaller information systems where no dedicated administrator is available to maintain the category scheme. This was demonstrated through a number of experiments evaluating these approaches.

10.2.1 *Contribution to Knowledge*

As an exploratory study, this thesis has covered a very broad range of research. This has included some very different fields such as Cognitive Science, Computer Science, LIS, and HCI. Each of these areas has played an important part, first in understanding the categorisation problem, and second in identifying potential solutions. Since this thesis has covered such a broad range of research, its original contributions are also quite varied. This section outlines contributions made by each chapter.

Chapter 2. Categorisation in a Manufacturing Environment.

Chapter 2 investigated what symptoms indicate that a categorisation problem exists; what does a problematic categorisation system look like? Through a case study of categorisation in SIMPRESS, two key indicators of categorisation problems were identified. Firstly, uneven category usage, and secondly, over use of an *Other* or *Miscellaneous* category.

In addition to symptoms, Chapter 2 also investigated the causes of categorisation problems identified in SIMPRESS. Interviews with plant staff and analysis of the system design revealed that categorisation problems are not solely the result of poor system design, nor just individual perceptions of the system, but the organisational context and culture also plays a significant part in determining how the system is used. This shows that the investigation of categorisation problems must necessarily be multidisciplinary to cover the broad range of factors that impact on categorisation.

Chapter 3. Cognitive Reasons for Categorisation Problems

Chapter 3 examined the role of categorisation in the mind through a review of the cognitive science literature. The review identified a number of cognitive factors that can cause categorisation problems in information systems. Based on these factors, the chapter draws implications for the design of information systems.

Chapter 4. Graded Categorisation and User Interfaces.

Chapter 3 investigated the impact of an interface allowing for graded category membership. Chapter 3 showed that categories in the mind have a graded structure, so it seemed reasonable to hypothesise that allowing graded membership would increase categorisation accuracy.

interfaces for categorisation. The results showed that using the graded interface increased accuracy by around 3%; however, people using the graded interface took, on average, 8.8 seconds longer to categorise. The results also showed that individual people not particularly good at predicting the categorisations of their peers, with pair-wise correlations of only 0.41–0.43. In spite of this, the data suggests that the collective contributions of users can be aggregated to counteract individual disagreement. Hence in systems with multiple users, collaborative categorisation may be a viable alternative to employing expert administrator.

Chapter 5. Contextual Reasons for Categorisation Problems.

Although cognitive factors are important, Chapter 2 showed that contextual factors have a significant impact on categorisation. Chapter 5 surveyed the literature on contextual causes, showing that the kinds of contextual issues observed with SIMPRESS were not unique to the organisation, but are common in many situations. The survey also revealed a gap in the literature (examined further in Chapter 6) concerning assumptions about how categorisation problems are to be solved.

Chapter 6. Potential Solutions.

Chapter 6 further clarified the gap in the literature, surveying various approaches that others have taken to solving categorisation issues. None of the approaches surveyed completely fulfilled all the requirements as set out in Chapter 5. Reviewing the strengths and weaknesses of each approach however, Chatper 6 identifies folksonomies as showing the most potential of being adapted to suit the requirements.

Chapter 7. Investigation of Folksonomy Tag Clouds.

While there has been much debate about tag clouds amongst information architects, very few have supported their claims with empirical evidence. Chapter 7 reported on the results of an experiment evaluating the use of tag clouds for information seeking. The results showed that the tag cloud is indeed a useful tool for browsing a corpus, but is not so useful for finding specific items. So, although tag clouds are not sufficient as the sole means of accessing information in an information system, they provide a valuable method for utilising folksonomy data.

Chapter 8. Comparison of Folksonomy Clustering Techniques.

A number of researchers have proposed clustering as a method of improving the usefulness of folksonomies. Clustering has the potential to disambiguate tags, identify social networks, and improve navigation. Each researcher however, proposes a different clustering method with little justification for their choice. This chapter reports on an experiment comparing a number of different clustering algorithms applied to a folksonomy dataset. The results showed that the ROCK algorithm performs best, but is too computationally expensive to be useful in most circumstances.

10.3 FUTURE WORK

This thesis has shown that folksonomies have great potential to improve categorisation in smaller information systems where there is no dedicated administrator available. Folksonomies are an exciting area of research with many important questions still unanswered. Some important areas for future investigation include:

TAG FEEDBACK MECHANISMS. One of the main drawbacks of folksonomies is the issue of homonymy. That is, where multiple words can be used to describe the same concept. This issue is aggravated in folksonomies by the possibility of inconsistent labelling. For example, one person may use the tag *webdesign*, another *web_design*, and still another *web design*. One method to combat this issue is to have the tagging interface 'suggest' tags that other people have used. The theory is that once people see how others have tagged, they are more likely to conform to the group standard. These suggestion mechanisms (in various guises) have been a feature of popular tagging services such as Delicious and RawSugar. However, little research has been published however on the effectiveness of this approach.

KNOWLEDGE SHARING. Because of the social nature of folksonomies, there is potential for tagging services to promote knowledge sharing within an organisation. The folksonomy identifies a person by the resources they collect, which may give a good indication of expertise in an area. Conversely, someone known to be an expert may add authority to the resources in their collection. While there have been a few preliminary studies into the use of folksonomies in an enterprise context, very few evaluate the

usefulness of the folksonomy as an aid to knowledge sharing.

FOLKSONOMY CLUSTERING. The clustering study (Chapter 8) reported that the ROCK algorithm produced superior clustering results, but did so at considerable computational cost. Reducing this computational cost without compromising cluster quality would be a useful area of investigation. Such an investigation may involve evaluating even more clustering approaches using ROCK as a baseline measurement. Alternatively, methods could be sought to reduce the complexity of the ROCK algorithm itself.

Part III

APPENDIX

A

CONCERN TYPES IN Simpress

Bedding	Light-On	Slug
Broken Finger	Location Pins	Slug Build-up
Broken Punch	Loose Metal	Splits
Broken Transfer Finger	Lows	Springback
Buckles	Margin	Steel Properties
Burrs	Measurement	Streamers
CAD	Metal Folds	Tear Outs
Closing Effort	Misalignment	Thinning
Contamination	Misfeed	Timing
Damage	Mislocation	Tolerance
Design	Pimples	Trim / Flange
Dunnage	Runs	Trim Edge
Facility	Safety	Variation
Flange Clearance	Scores	Weakness
Fouling	Scrap Build-up	Weld Integrity
Gripper	Sensor	Weld Quality
Highs	Slip Planes	Wrinkles
Lifter	Slivers	

B

POPCAT INFORMATION SHEET

This is the information sheet handed out to participants in the experiment described in Chapter 4.

B.1 CATEGORISATION

Categorisation is one of the most fundamental processes of the human brain. According to Lakoff [70], "without the ability to categorize, we could not function at all, either in the physical world or in our social and intellectual lives." Yet, when entering data into an information system, humans are often required to categorise in a way which does not reflect the natural processes of the mind. Furthermore, much energy is spent on attempting to make computers categorise bodies of text for us, when human beings are inherently adept at the task.

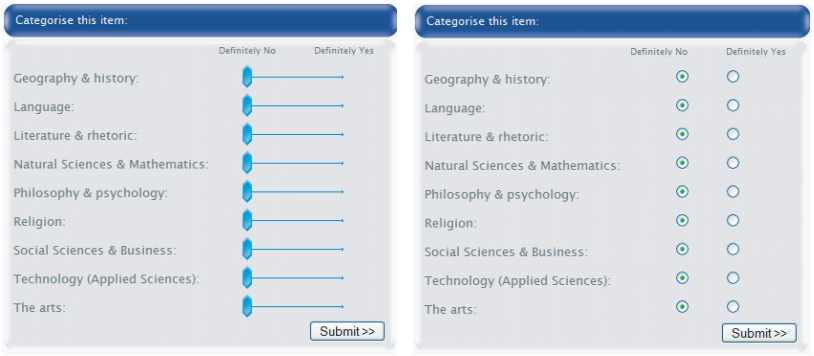
The purpose of the research is to investigate ways that user-interfaces can be made to more naturally support human categorisation processes. It is hoped that this will lead to decreased cognitive load for the user and better organised data. The first stage of the study, PopCat, is examining the use of graded membership in categorisation processes.

B.2 THE STUDY

The PopCat study aims to compare two different user interfaces to see how they affect categorisations made by people. One interface allows users to say only that an item *does*, or *does not* belong to a particular category.

The other interface allows users to say that an item *sort of* belongs to a category or belongs *a little bit*, by indicating degree of membership using a slider bar.

The interfaces will be compared by asking participants to play a game. In the game, players are shown an item such as an abstract from an academic journal, and are asked to indicate whether it belongs to certain categories. The score for each round is based on how closely the player's answers matched those of other players, i.e. players whose answers are closest to the mean get the highest scores.



(a) (b)

Figure 51: The two categorisation interfaces: (a) The slider-bar interface; (b) The radio button interface.

Different groups of users will play the game using the two interfaces. The results will be compared to see whether one interface produces a greater consensus of categorisations than the other. The time taken to make categorisations will also be compared between the two interfaces.

B.3 HOW DO I PLAY?

The game is normally played in groups of four to eight people. To play, you first need to *sign up* and create a user account. Once you have done that, you can click on the "play" link. You will be given the choice of creating a new game or joining an existing one. To join a game, just click on the username of the person who created it. Once you have joined, you will have to wait until the person who created it clicks the "start game" button on their screen.

Playing the game is simple. You will be shown an item such as a picture, a piece of text or a video. You will then be asked to decide whether or not the item belongs to some different categories. The idea is to try and think about how other people would categorise the item because your score is based on what everyone in the your group picks. The closer you are to the mean choices for that round, the higher the score you receive.

C

SLASHDOT CATEGORY FREQUENCIES

Category	Frequency	Category	Frequency
Science	473	Media	3
Space	196	Sci-Fi	3
Technology	113	Quickies	3
News	81	Software	3
Biotech	79	Programming	2
NASA	42	Entertainment	2
Power	25	Interviews	2
Politics	23	IBM	2
Robotics	19	Book Reviews	2
Education	18	The Media	2
Math	12	The Courts	2
The Almighty Buck	11	Slashback	2
Mars	11	Censorship	2
Hardware	9	Encryption	2
Businesses	9	Graphics	2
United States	8	Supercomputing	2
Index	8	Bug	1
Communications	7	Wireless Networking	1
Games	7	Worms	1
The Internet	7	Books	1
Ask Slashdot	6	Announcements	1
Moon	6	Anime	1
Google	6	XBox (Games)	1
It's funny. Laugh.	6	Movies	1
Editorial	6	Networking	1
Privacy	5	Input Devices	1
Television	5	Databases	1
Patents	5	Linux	1
Security	4	Hardware Hacking	1
IT	3	Star Wars Prequels	1
Toys	3	First Person Shooters (Games)	1

D

DATASET TAG FREQUENCIES

KEYWORD	FREQ.	KEYWORD	FREQ.	KEYWORD	FREQ.
NASA	68	Biodiesel	3	silicon	2
Space	62	exploration	3	new york times	2
Science	53	mathematics	3	energy systems	2
mars	18	probe	3	science fiction	2
biology	14	Milky Way	3	Suitsat	2
global warming	13	saturn	3	deep impact	2
health	13	hybrid	3	transmitter	2
Moon	13	organic	3	materials	2
physics	12	vehicle	3	original manuscript	2
Greenhouse	11	gas	3	grammar	2
DNA	11	satellites	3	Hemos	2
Computing	11	SpaceShipOne	3	Flights	2
Astronomy	10	cassini	3	wine	2
robot	9	microsoft	3	rats	2
cancer	9	Cloning	3	solar system	2
evolution	9	Optics	3	astrophysics	2
MIT	8	Games	3	cnn	2
telescope	8	storm	3	flight	2
fusion	8	boeing	3	rings	2
virus	7	satelight	3	vehicles	2
Brett Favre	7	stem cells	3	internet	2
brain	7	climate change	3	military	2
Satellite	7	nerve	3	debate	2
technology	7	algae	3	Solar Power	2
climate	7	ocean	3	shuttle	2
electricity	7	human	3	Death	2
bacteria	6	water	3	astronauts	2
rocket	6	leonardo da vinci	3	LiveScience	2
things muhummad ali doesn't like	6	education	3	Huwmerous articles	2
Green	6	japan	3	microwaves	2

KEYWORD	FREQ.	KEYWORD	FREQ.	KEYWORD	FREQ.
SpaceX	6	Hubble Space Telescope	3	String theory	2
International Space Station	6	aids	3	wild 2	2
DARPA	6	frequency	2	temperature	2
galaxy	6	ai	2	Solar Flares	2
spacecraft	6	apollo	2	orbit	2
research	6	falcon 1	2	stress	2
Medicine	5	space junk	2	alzheimers	2
ice	5	Aging	2	gaming	2
genetics	5	Bush Administration	2	microscopy	2
HIV	5	Money	2	solar	2
history	5	earthquake	2	bones	2
robotics	5	melting	2	fish	2
genes	5	learning	2	hurricane	2
energy	5	birds	2	emotion	2
Venus	5	nuclear	2	microscope	2
Google	5	Religion	2	carpal tunnel	2
quantum	5	space filght	2	game addiction	2
India	5	antioxidants	2	radiation	2
stars	5	Face transplant	2	2007	2
discovery	5	Voyager	2	asia	2
Russia	5	model	2	optical	2
China	5	bird	2	galileo	2
stardust	5	magnetic field	2	white house	2
hubble	5	particle	2	linux	2
communications	4	mobile phones	2	Humans	2
nuclear fusion	4	Aerospace	2	phone	2
nanotubes	4	medicin	2	god	2
space elevator	4	ethics	2	electronics	2
ScuttleMonkey	4	scramjet	2	insects	2
earth	4	psychology	2	marketing	2
nanotechnology	4	fossils	2	US	2
ESA	4	genome	2	carbon dioxide	2
cells	4	lasers	2	insect	2
Coffee	4	ISS	2	stem cell	2
gravity	4	comets	2	power	2
Hydrogen	4	scientists	2	Uranus	2

KEYWORD	FREQ.	KEYWORD	FREQ.	KEYWORD	FREQ.
GPS	4	eyes	2	carbon	2
enceladus	4	holograms	2	electronic	2
big bang	4	ISRO	2	Health Care	2
asteroid	4	comet	2	LiftPort	2
intelligent design	4	space agency	2	future	2
neuroscience	4	plant	2	SALT	2
dark matter	4	viruses	2	army	2
Light	4	anthrax	2	smoking	2
Aliens	4	Supersonic Jet	2	blood vessels	2
life	4	things muhummad ali respects	2	quantum physics	2
Einstein	4	pandemic	2	Car	2
disease	4	Antarctica	2	damage	2
space shuttle	4	planets	2	information	2
fuel	4	Avation	2	display	2
black hole	4	eye	2	replacement	2
dinosaur	4	soyuz	2	influenza	2
sleep	3	mistakes	2	weapons	2
patent	3	Nanomachines	2	alcohol	2
chemistry	3	protection	2	methane	2
Stanford	3	Dementia	2	blindness	2
time	3	IQ	2	blind	2
Africa	3	Autism	2	Space Station	2
bird flu	3	lander	2	university of kansas	2
pluto	3	animals	2	Archeology	2
geology	3	Mission	2	universe	2
biotech	3	record	2	archaeology	2
IT	3	Spider	2	computer	2
nanotech	3	Green energy	2	cyborg	2
launch	3	genetic engineering	2	FDA	2
addiction	3	engineers	2	japanese	2
language	3	hyperlink example of	2	Police	2
proteins	3	maths	2	FBI	2
robots	3	video games	2	aircraft	2
environment	3	engineering	2	children	2
supernova	3	atomic	2	Titan	2
Nobel Prize	3	travel	2	warming	2
planet	3	cell phone	2	Rover	2
lunar	3	ancient	2	ultrasound	2

KEYWORD	FREQ.	KEYWORD	FREQ.	KEYWORD	FREQ.
nano	3	engine	2		

E

CLUSTERING RESULTS

Listed here are the 10 largest clusters that each clustering algorithm produced for the SlashDot dataset. Each algorithm was run to produce the optimal number of clusters, according to the Q-measure. The list shows tags and categories in the cluster that occurred more than once. The frequency of each tag and category in the cluster is shown in brackets.

K-Means

256 Clusters, 137 single-item clusters

SIZE	TAGS	CATEGORIES
37	space (36) astronomy (3) satelight (3) future (2) china (2) exploration (2) russia (2) communications (2) milky way (2)	Space (36) Science (17) News (7) Technology (4) NASA (3) Index (2) Businesses (2)
25	science (24) physics (3) geology (2) optics (2)	Science (23) Technology (6) News (4) Biotech (3) Editorial (2) Sci-Fi (2)
16	nasa (16) spacecraft (2)	Space (11) Science (9) NASA (5) Technology (3)
13	global warming (13) greenhouse (5) climate change (2)	Science (12) News (2)
12	science (11) space (9) nasa (6) comets (2)	Science (9) Space (9) NASA (6) Politics (2)
10	health (9) antioxidants (2)	Science (8) The Internet (2)

	coffee (2) computing (2)	Biotech (2)
8	nasa (7) space (6) hubble space telescope (2)	Space (7) Science (5) NASA (3) News (2)
8	robot (8) darpa (2)	Robotics (8) Science (8)
8	mars (8)	Mars (4) Space (4) News (2) Science (2)
8	evolution (8)	Science (8) Biotech (3) News (2)

Agglomerative

286 clusters, 173 Single-entry clusters

SIZE	TAGS	CATEGORIES
48	space (48) nasa (15) science (9) astronomy (3) satelight (3) international space station (2) china (2) space shuttle (2) hubble (2) communications (2) esa (2) hubble space telescope (2) exploration (2)	Space (44) Science (28) NASA (12) News (8) Technology (4) Index (3) Businesses (3) Politics (2)
35	nasa (35) mars (6) spacecraft (3) science (2) discovery (2)	Space (21) Science (20) NASA (12) Mars (5) Technology (4) News (2) Politics (2)
33	science (33) biology (7) physics (4) dna (3)	Science (31) Technology (8) Biotech (7) News (6)

	optics (2) computing (2) engineering (2) materials (2)	Editorial (2) Privacy (2) Education (2) Power (2) Politics (2)
11	global warming (11) greenhouse (5) climate change (2)	Science (10) News (2)
8	health (8) computing (3) biology (2)	Science (7) Biotech (3) Games (2) It's funny. Laugh. (2)
8	evolution (8)	Science (8) Biotech (3) News (2)
7	moon (7) nasa (2)	Science (4) Space (4) NASA (3) Moon (2)
7	robot (7)	Robotics (7) Science (7)
6	fusion (6) electricity (2)	Science (5) Power (3)
6	galaxy (6) milky way (2)	Space (6) Science (4)

ROCK

286 Clusters, 228 single-entry clusters.

SIZE	TAGS	CATEGORIES
105	nasa (62) space (60) science (11) mars (7) moon (4) exploration (3) satelight (3) astronomy (3) esa (3) international space station (3) russia (3) hubble (3) spacecraft (3)	Space (82) Science (62) NASA (29) News (11) Technology (10) Mars (6) Politics (6) Index (4) Businesses (3) The Almighty Buck (3) Education (2) Moon (2) Robotics (2)

SIZE	TAGS	CATEGORIES
	hubble space telescope (2) space shuttle (2) communications (2) china (2) astronauts (2) lunar (2) soyuz (2) robotics (2) discovery (2) planets (2) spacex (2)	
47	science (41) biology (14) physics (4) genetics (3) dna (3) engineering (2) materials (2) computing (2) geology (2) optics (2) health (2)	Science (44) Biotech (11) Technology (10) News (9) Editorial (3) Sci-Fi (2) Privacy (2) Education (2) Power (2) Politics (2)
17	global warming (11) greenhouse (10) climate change (3) climate (2)	Science (14) News (3) Space (3) Biotech (2) Politics (2)
16	health (11) computing (8) coffee (2) health care (2)	Science (13) Biotech (3) The Internet (2) Games (2) It's funny. Laugh. (2) Hardware (2) Technology (2)
11	physics (7) astronomy (6)	Science (9) Space (7) News (2)
10	mars (10) nasa (2)	Space (6) Mars (4) News (2) NASA (2) Science (2)
8	robot (8)	Robotics (8)

SIZE	TAGS	CATEGORIES
	darpa (2)	Science (8)
8	evolution (8)	Science (8) Biotech (3) News (2)
7	brett favre (5) things muhummad ali doesn't like (4)	Science (3) Technology (2) Space (2)
6	dna (6)	Science (5) Biotech (3)

CLOPE

220 clusters, 81 single-entry clusters.

SIZE	TAGS	CATEGORIES
28	space (11) nasa (9) science (4) india (4) spacecraft (4) discovery (3) space shuttle (3) esa (3) exploration (3) international space station (3) russia (3) moon (3) spacex (3) rocket (2) hubble space telescope (2) space station (2) comets (2) planets (2) venus (2) soyuz (2) isro (2)	Space (25) Science (15) Technology (6) NASA (3) Politics (3) News (2)
20	nasa (12) space (11) science (7) mars (3) moon (3) international space station (2)	Space (14) Science (11) NASA (9) News (2) Politics (2)
16	nasa (9)	Science (9)

	mars (7)	Space (7)
	science (4)	Mars (5)
	geology (3)	NASA (4)
	life (2)	The Almighty Buck (2)
	space junk (2)	
14	space (13)	Space (13)
	nasa (7)	Science (8)
	satelight (3)	News (4)
	astronomy (3)	NASA (3)
	communications (2)	
	hubble (2)	
14	computing (8)	Science (12)
	science (8)	Biotech (3)
	health (3)	Technology (3)
	green energy (2)	Games (2)
	nanomachines (2)	Education (2)
	health care (2)	
10	electricity (4)	Science (7)
	fusion (4)	Space (4)
	space (3)	Technology (3)

[1] Aamodt, A. and Plaza, E. [1994], 'Case-based reasoning: foundational issues, methodological variations, and system approaches', *AI Communications* **7**(1), 39–59. (Cited on page 73.)

[2] Abbott, R. [2004], 'Subjectivity as a concern for information science: a Popperian perspective', *Journal of Information Science* **30**(2), 95–106. (Cited on page 76.)

[3] Albrechtsen, H. [2000], Who wants yesterday's classifications?—information science perspectives on classification schemes in common information spaces, *in* K. Schmidt, ed., 'Workshop on 'Cooperative Organization of Common Information Spaces', Lyngby, Denmark, 23–25 August 2000'. Retrieved 6 June 2007. **URL:** *http://www.itu.dk/people/schmidt/ciscph2000/Albrechtsen.pdf* (Cited on pages 86 and 87.)

[4] Albrechtsen, H. and Jacob, E. K. [1998], 'The dynamics of classification systems as boundary objects for cooperation in the electronic library', *Library Trends* **47**(2), 293–312. (Cited on page 72.)

[5] Albrechtsen, H. and Pejtersen, A. M. [2003], 'Cognitive work analysis and work centered design of classification schemes', *Knowledge Organization* **30**(3/4), 313–227. (Cited on pages 86 and 87.)

[6] Albrechtsen, H., Pejtersen, A. M. and Cleal, B. [2002], Empirical work analysis of collaborative film indexing, *in* H. Bruce, R. Fidel, P. Ingwersen and P. Vakkari, eds, 'Emerging Frameworks and Methods: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS4)', Libraries Unlimited, Greenwood Village, CO, USA, pp. 85–107. (Cited on pages 86 and 87.)

[7] Anderson, J. D. and Pérez-Carballo, J. [2001*a*], 'The nature of indexing: how humans and machines analyze messages and texts for retrieval: part i: research, and the nature of human indexing', *Information Processing & Management* **37**(2), 231–254. (Cited on pages 108 and 109.)

[8] Anderson, J. D. and Pérez-Carballo, J. [2001*b*], 'The nature of indexing: how humans and machines analyze messages and texts for retrieval: part ii: machine indexing, and the allocation of human versus machine effort', *Information Processing & Management* **37**(2), 255–277. (Cited on page 108.)

[9] Barnum, C., Henderson, E., Hood, A. and Jordan, R. [2004], 'Index versus full-text search: A usability study of user preference

and performance', *Technical Communication* **51**(2), 185–206. (Cited on page 109.)

[10] Barsalou, L. W. [1983], 'Ad hoc categories', *Memory & Cognition* **11**(3), 211–227. (Cited on pages 33, 40, and 48.)

[11] Barsalou, L. W. [1987], The instability of graded structure: implications for the nature of concepts, *in* U. Neisser, ed., 'Concepts and conceptual development: The ecological and intellectual factors in categorisation', Cambridge University Press, pp. 101–140. (Cited on page 61.)

[12] Barton, J., Currier, S. and Hey, J. M. N. [2003], Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice, *in* S. Sutton, J. Greenberg and J. Tennis, eds, 'DC-2003: Proceedings of the International DCMI Metadata Conference and Workshop.', Information Institute of Syracuse, Syracuse, NY, USA. (Cited on pages 110 and 111.)

[13] Begelman, G., Keller, P. and Smadja, F. [2006], Automated tag clustering: Improving search and exploration in the tag space, *in* 'Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland'. Retrieved 3 July 2006.
URL: *http://www.rawsugar.com/www2006/automatedTagClustering.pdf*
(Cited on pages 148 and 149.)

[14] Bielenberg, K. and Zacher, M. [2005], Groups in social software: Utilizing tagging to integrate individual contexts for social navigation, Master's thesis, Universität Bremen. (Cited on pages 150, 151, 175, 176, and 195.)

[15] Bowker, G. C. and Star, S. L. [1991], Situations vs. standards in long-term, wide-scale decision-making: The case of the international classification of diseases, *in* J. F. Nunamaker, Jr. and R. H. Sprague, Jr., eds, 'Proceedings of the Twenty-Fourth Annual Hawaii International Conference on System Sciences', Vol. iv, IEEE Computer Society Press, pp. 73–81. (Cited on pages 72 and 73.)

[16] Bowker, G. C. and Star, S. L. [1999], *Sorting Things Out: Classification and Its Consequences*, Inside Technology Series, The MIT Press, Cambridge, Massachusetts. (Cited on pages 1, 7, 23, 72, 73, 74, 75, and 77.)

[17] Brin, S. and Page, L. [1998], 'The anatomy of a large-scale hypertextual web search engine', *Computer Networks and ISDN Systems* **30**(1-7), 107–117. (Cited on page 110.)

[18] Brooks, C. H. and Montanez, N. [2006], Improved annotation of the blogosphere via autotagging and hierarchical clustering, *in* 'WWW '06: Proceedings of the 15th international conference on

World Wide Web', ACM Press, pp. 625–632. (Cited on pages 7, 126, 134, and 148.)

[19] Broughton, V. [2006], 'The need for a faceted classification as the basis of all methods of information retrieval', *Aslib proceedings : New information perspectives* **56**(1/2), 49–72. (Cited on pages 89 and 92.)

[20] Brown, J. S., Collins, A. and Duguid, P. [1989], 'Situated cognition and the culture of learning', *Educational Researcher* **18**(1), 32–42. (Cited on pages 39 and 86.)

[21] Brown, R. [1958], 'How shall a thing be called?', *Psychological Review* **65**(1), 14–21. (Cited on pages 34, 42, and 71.)

[22] Buckland, M. [1999], Vocabulary as a central concept in library and information science, *in* T. Aparac, T. Saracevic, P. Ingwersen and P. Vakkari, eds, 'Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities. Proceedings of the Third International Conference on Conceptions of Library and Information Science (CoLIS3)', Benja Publishing, Lokve, Croatia. (Cited on pages 42 and 81.)

[23] Cardew-Hall, M., Gresham, R. J., Hodgson, P. and Smith, J. I. [2000], Knowledge capture system for sheet metal forming problems, *in* 'Eighth International Conference On Manufacturing Engineering (ICME2000)'. (Cited on pages 14 and 15.)

[24] Cardew-Hall, M., Smith, J. I., Pantano, V. and Hodgson, P. [2002], Knowledge capture and feedback system for design and manufacture, *in* 'SAE International Body Engineering Conference - IBEC 2002, Paris, France', SAE. (Cited on page 15.)

[25] Cattuto, C., Loreto, V. and Pietronero, L. [2006], 'Collaborative tagging and semiotic dynamics'. Arxiv preprint cs.CY/0605015. Retrieved 6 May 2006.
URL: *http://arxiv.org/abs/cs.CY/0605015* (Cited on pages 126 and 128.)

[26] Cockburn, A. [2000], *Writing Effective Use Cases*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. (Cited on page 176.)

[27] Collantes, L. Y. [1995], 'Degree of agreement in naming objects and concepts for information retrieval', *Journal of the American Society for Information Science* **46**(2), 116–132. (Cited on pages 43, 45, 52, and 81.)

[28] Cutting, D. R., Pedersen, J. O., Karger, D. and Tukey, J. W. [1992], Scatter/gather: A cluster-based approach to browsing large document collections, *in* 'Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', pp. 318–329. (Cited on pages 148 and 161.)

[29] Damianos, L., Griffith, J., Cuomo, D., Hirst, D. and Smallwood, J. [2006], Onomi: Social bookmarking on a corporate intranet, *in* 'Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland'. Retrieved 3 July 2006.
URL: *http://www.topixa.com/www2006/28.pdf* (Cited on page 132.)

[30] de Mántaras, R. L. and Plaza, E. [1997], 'Case-based reasoning:: an overview', *AI Communications* **10**(1), 21–29. (Cited on page 73.)

[31] Denton, W. [2003], 'How to make a faceted classification and put it on the web'. Last updated 15 February 2007. Retrieved 7 June 2007.
URL: *http://www.miskatonic.org/library/facet-web-howto.html* (Cited on pages 90 and 92.)

[32] Dieberger, A., Dourish, P., Höök, K., Resnick, P. and Wexelblat, A. [2000], 'Social navigation: Techniques for building more usable systems', *Interactions* **7**(6), 36–45. (Cited on page 124.)

[33] Doctorrow, C. [2001], 'Metacrap: Putting the torch to seven strawmen of the meta-utopia'. Retrieved 27 October 2006.
URL: *http://www.well.com/~doctorow/metacrap.htm* (Cited on pages 48 and 111.)

[34] Edmunds, A. and Morris, A. [2000], 'The problem of information overload in business organisations: a review of the literature', *International Journal of Information Management* **20**(1), 17–28. (Cited on page 3.)

[35] Estes, W. K. [1994], *Classification and Cognition*, Oxford University Press. (Cited on page 1.)

[36] Farrell, S. and Lau, T. [2006], Fringe contacts: People-tagging for the enterprise, *in* 'Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, 22-26 May 2006'. Retrieved 3 July 2006.
URL: *http://www.rawsugar.com/www2006/25.pdf* (Cited on page 132.)

[37] Farrow, J. [1995], 'Academic indexing: what's it all about?', *The Indexer* **19**(4), 243–247. (Cited on page 39.)

[38] Fidel, R. [1992], 'Who needs controlled vocabulary?', *Special Libraries* **83**(1), 1–9. (Cited on pages 108 and 109.)

[39] Fidel, R. and Pejtersen, A. M. [2004], 'From information behaviour research to the design of information systems: the cognitive work analysis framework', *Information Research* **10**(1). Paper 210. Retrieved 25 May 2007.
URL: *http://InformationR.net/ir/10-1/paper201.html* (Cited on pages 85, 86, and 87.)

[40] Fidel, R., Pejtersen, A. M., Cleal, B. and Bruce, H. [2004], 'Multidimensional approach to the study of human–information interaction: A case study of collaborative information retrieval', *Journal of the American Society for Information Science and Technology* **55**(11), 939–953. (Cited on pages 85, 86, 87, and 88.)

[41] Fidelman, E. [2006], Metadata quality and the use of hierarchical schemes to determine meta keywords: An exploration, Master's thesis, School of Information and Library Science, University of North Carolina, Chapel Hill. Retrieved 7 June 2007.
URL: *http://hdl.handle.net/1901/288* (Cited on pages 48 and 89.)

[42] Furnas, G. W., Landauer, T. K., Gomez., L. M. and Dumais, S. T. [1987], 'The vocabulary problem in human–system communication', *Communications of the ACM* **30**(11), 964–971. (Cited on pages 43, 45, and 81.)

[43] Gardner, H. [1985], *The Mind's New Science: A History of the Cognitive Revolution*, Basic Books, New York, NY, USA. (Cited on page 32.)

[44] Georgakopoulos, D., Hornick, M. and Sheth, A. [1995], 'An overview of workflow management: From process modelling to workflow automation infrascructure', *Distributed and Parallel Databases* **3**(2), 119–153. (Cited on page 73.)

[45] Golder, S. A. and Huberman, B. A. [2006], 'Usage patterns of collaborative tagging systems', *Journal of Information Science* **32**(2), 198–208. (Cited on page 126.)

[46] Greenberg, J., Pattuelli, M. C., Parsia, B. and Robertsin, W. D. [2001], 'Author-generated dublin core metadata for web resources: A baseline study in an organization', *Journal of Digital Information* **2**(2). (Cited on page 111.)

[47] Gross, T. and Taylor, A. G. [2005], 'What have we got to lose? the effect of controlled vocabulary on keyword searching results', *College and Research Libraries* **66**(3), 212–230. (Cited on page 109.)

[48] Guha, S., Rastogi, R. and Shim, K. [1997], A clustering algorithm for categorical attributes, Technical report, Bell Laboratories, Murray Hill, USA.
URL: *http://citeseer.ist.psu.edu/guha97clustering.html* (Cited on page 163.)

[49] Guha, S., Rastogi, R. and Shim, K. [1999], ROCK: A robust clustering algorithm for categorical attributes, *in* '15th International Conference on Data Engineering (ICDE'99)', IEEE Computer Society. (Cited on page 162.)

[50] Hammond, T., Hannay, T., Lund, B. and Scott, J. [2005], 'Social bookmarking tools (i) a general review', *D-Lib Magazine* **11**(4). Retrived 7 October 2006.

URL: *http://dx.doi.org/10.1045/april2005-hammond* (Cited on pages 111 and 121.)

[51] Hampton, J. A. [1979], 'Polymorphous concepts in semantic memory', *Journal of Verbal Learning and Verbal Behavior* **18**, 441–461. (Cited on page 33.)

[52] Hanson, F. A. [2004], 'From classification to indexing: How automation transforms the way we think', *Social Epistemology* **18**(4), 333–356. (Cited on page 98.)

[53] Henry, J. W. and Stone, R. W. [1995], 'Computer self-efficacy and outcome expectancy: the effects on the end-user's job satisfaction', *SIGCPR Comput. Pers.* **16**(4), 15–34. (Cited on page 78.)

[54] Henziger, M. R. [2001], 'Hyperlink analysis for the web', *IEEE Internet Computing* **5**(1), 45–50. (Cited on page 110.)

[55] Hjørland, B. [1998*a*], 'The classification of psychology: A case study in the classification of a knowledge field', *Knowledge Organization* **25**(4), 162–201. (Cited on page 86.)

[56] Hjørland, B. [1998*b*], 'Theory and metatheory of information science: A new interpretation', *Journal of Documentation* **54**(5), 606–621.

[57] Hjørland, B. [2002], 'Domain analysis in information science: Eleven approaches—traditional as well as innovative', *Journal of Documentation* **58**(4), 422–462. (Cited on page 87.)

[58] Hjørland, B. and Albrechtsen, H. [1995], 'Towards a new horizon in information science: Domain-analysis', *Journal of the American Society for Information Science* **46**(6), 400–425. (Cited on pages 85 and 86.)

[59] Hjørland, B. and Albrechtsen, H. [1999], 'An analysis of some trends in classification research', *Knowledge Organization* **26**(3), 131–139. (Cited on pages 74 and 81.)

[60] Hotho, A., Jäschke, R., Schmitz, C. and Stumme, G. [2006], Information retrieval in folksonomies: search and ranking, *in* Y. Sure and J. Domingue, eds, 'The Semantic Web: Research and Applications, 3rd European Semantic Web Conference, ESWC 2006, Budva, Montenegro, 11–14 June 2006', Springer, Heidelberg, pp. 411–426. (Cited on page 131.)

[61] Huysman, M. and Wulf, V. [2006], 'It to support knowledge sharing in communities, towards a social capital analysis', *Journal of Information Technology* **21**(1), 40. (Cited on page 124.)

[62] Jacob, E. K. [2004], 'Classification and categorization: A difference that makes a difference', *Library Trends* **52**(3), 515–540. (Cited on pages 7, 31, 32, 35, 36, 38, 40, and 71.)

[63] Jacobsen, D. [1995], 'Field list for the canonical bibliography format'. Last updated 17 January 1995. Retrieved 30 September 2005.
URL: *http://www.ecst.csuchico.edu/~jacobsd/bib/bp/CanonicalFields.html* (Cited on page 185.)

[64] Jain, A. K., Murty, M. N. and Flynn, P. J. [1999], 'Data clustering: a review', *ACM Computing Surveys* **31**(3), 264–323. (Cited on pages 144, 145, and 151.)

[65] Jon, N. [2005], 'Tag clouds: A response'. Retrieved 18 July 2006.
URL: *http://www.nicholasjon.com/archives/2005/5/4/tag_clouds_a_response* (Cited on page 125.)

[66] Kroski, E. [2005], 'The hive mind: Folksonomies and user-based tagging'. Retrieved 29 January 2007.
URL: *http://infotangle.blogsome.com/2005/12/07/the-hive-mind-folksonomies-and-user-based-tagging/* (Cited on page 112.)

[67] Kwasnik, B. H. [1999], 'The role of classification in knowledge representation and discovery', *Library Trends* **48**(1), 22–47. (Cited on pages 72 and 90.)

[68] La Barre, K. [2004], Adventures in faceted classification: A brave new world or a world of confusion?, *in* I. C. McIlwaine, ed., 'Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference, 13-16 July 2004, London, UK', Vol. 9 of *Advances in Knowledge Organization*, Ergon-Verlag, Würtzburg, Germany, pp. 79–84. (Cited on pages 89 and 90.)

[69] La Barre, K. [2007], The Use of Faceted Analytico-Synthetic Theory as Revealed in the Practice of Website Construction and Design, PhD thesis, Indiana University, Bloomington, USA. Retrieved 7 June 2007.
URL: *http://leep.lis.uiuc.edu/publish/klabarre/dissBYsection.html* (Cited on pages 89 and 90.)

[70] Lakoff, G. [1987], *Women, Fire, and Dangerous Things*, University of Chicago Press, Chicago. (Cited on pages 1, 7, 70, and 211.)

[71] Lakoff, G. and Johnson, M. [1980], *Metaphors We Live By*, University of Chicago Press, Chicago. 2003 reprint with new afterword. (Cited on page 30.)

[72] Lakoff, G. and Johnson, M. [1999], *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*, Basic Books, New York. (Cited on pages 30, 31, and 35.)

[73] Lambiotte, R. and Ausloos, M. [2005], 'Collaborative tagging as a tripartite network'. Arxiv preprint cs.DS/0512090.
URL: *http://arxiv.org/abs/cs.DS/0512090* (Cited on pages 126 and 149.)

[74] Lambiotte, R. and Ausloos, M. [2006], Collaborative tagging as a tripartite network, *in* V. N. Alexandrov, G. D. van Albada, P. M. A. Sloot and J. Dongarra, eds, 'ICCS 2006: 6th International Conference on Computational Science', Vol. 3993 of *Lecture Notes in Computer Science*, Springer, Berlin, pp. 1114–1117. (Cited on pages 149, 150, and 184.)

[75] Larsen, B. and Aone, C. [1999], Fast and effective text mining using linear-time document clustering, *in* 'KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining', ACM Press, pp. 16–22. (Cited on page 155.)

[76] Lubbes, R. K. [2001], 'Automatic categorization: How it works, related issues, and impacts on records management', *Information Management Journal* **35**(4), 38–43. (Cited on pages 98 and 106.)

[77] Lubbes, R. K. [2003], 'So you want to implement automatic categorization?', *Information Management Journal* **37**(2), 60–69. (Cited on pages 95 and 104.)

[78] Maguire, M. [2001], 'Context of use within usability activities', *International Journal Human–Computer Studies* **55**(4), 453–483. (Cited on page 75.)

[79] Mai, J.-E. [2004], 'Classification in context: Relativity, reality, and representation', *Knowledge Organization* **31**(1), 39–48. (Cited on pages 71, 77, 78, 79, 81, 85, and 86.)

[80] Malone, T. W. [1983], 'How do people organize their desks?: Implications for the design of office information systems', *ACM Transactions on Information Systems (TOIS)* **1**(1), 99–112. (Cited on pages 2, 40, and 41.)

[81] Malt, B. C., Sloman, S. A., Gennari, S., Shi, M. and Wang, Y. [1999], 'Knowing versus naming: Similarity and the linguistic categorization of artifacts', *Journal of Memory and Language* **40**(2), 230–262. (Cited on pages 42, 52, and 70.)

[82] Mander, R., Salomon, G. and Wong, Y. Y. [1992], A "pile" metaphor for supporting casual organization of information, *in* 'CHI '92: Proceedings of the SIGCHI conference on Human factors in computing systems', ACM Press, New York, NY, US, pp. 627–634. (Cited on page 40.)

[83] Marchionini, G. [1995], *Information Seeking in Electronic Environments*, Cambridge Series on Human–Computer Interaction, Cambridge University Press. (Cited on pages 3, 112, and 138.)

[84] Markman, A. B. and Ross, B. H. [2003], 'Category use and category learning', *Psychological Bulletin* **129**(4), 592–613. (Cited on pages 37, 39, and 70.)

[85] Marlow, C., Naaman, M., Boyd, D. and Davis, M. [2006], HT06, tagging paper, taxonomy, flickr, academic article, to read, *in* U. K. Wiil, P. J. Nürnberg and J. Rubart, eds, 'Proceedings of the seventeenth conference on Hypertext and hypermedia, Odense, Denmark, 22–25 August 2006', ACM Press, New York, NY, USA, pp. 31–39. (Cited on page 127.)

[86] Mathes, A. [2004], 'Folksonomies - cooperative classification and communication through shared metadata', Doctoral Presentation, Graduate School of Library and Information Science, University of Illinois Urbana-Champaign. Retrieved 10 March 2006.
URL: *http://adammathes.com/academic/computer-mediated-communication/folksonomies.html* (Cited on pages 112, 113, 121, 126, and 134.)

[87] Maurer, D. and Warfel, T. [2007], 'Card sorting: a definitive guide'. Retrieved 23 May 2007.
URL: *http://boxesandarrows.com/view/card_sorting_a_definitive_guide* (Cited on pages 93 and 95.)

[88] McCloskey, M. E. and Glucksberg, S. [1978], 'Natural categories: Well defined or fuzzy sets?', *Memory and Cognition* **6**(4), 462–472. (Cited on page 34.)

[89] McGarty, C. [1999], *Categorization in Social Psychology*, SAGE Publications, London. (Cited on page 2.)

[90] McGeorge, P. and Rugg, G. [1992], 'The uses of 'contrived' knowledge elicitation techniques', **9**(3), 149–154. (Cited on page 93.)

[91] Merholz, P. [2004*a*], 'Ethnoclassification and vernacular vocabularies'. Accessed 29 January 2007.
URL: *http://www.peterme.com/archives/000387.html* (Cited on page 113.)

[92] Merholz, P. [2004*b*], 'Metadata for the masses'. Retrieved 3 January 2007.
URL: *http://www.adaptivepath.com/publications/essays/archives/000361.php* (Cited on page 113.)

[93] Merrell, F. [1995], *Semiosis in the Postmodern Age*, Purdue University Press, West Lafayette, USA. (Cited on page 79.)

[94] Mervis, C. B. and Rosch, E. [1981], 'Categorization of natural objects', *Annual Review of Psychology* **32**, 89–115. (Cited on page 34.)

[95] Meyers, J. [2002], 'Automatic categorization, taxonomies, and the world of information: Can't live with them, can't live without them', *E-Docs* **16**(6), 20–21. (Cited on pages 106 and 107.)

[96] Millen, D., Feinberg, J. and Kerr, B. [2005], 'Social bookmarking in the enterprise', *Queue* **3**(9), 28–35. (Cited on pages 113, 132, and 191.)

[97] Millen, D. R., Feinberg, J. and Kerr, B. [2006], Dogear: Social bookmarking in the enterprise, *in* R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries and G. Olson, eds, 'CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems', ACM Press, New York, NY, USA, pp. 111–120. (Cited on page 191.)

[98] Miller, P. [1996], 'Metadata for the masses', *Ariadne* **5**. (Cited on page 110.)

[99] Morville, P. [2006], *Information Architecture for the World Wide Web*, 3rd edn, O'Reilly Media, Sebastopol, CA, USA. (Cited on pages 44, 89, and 90.)

[100] Nielsen, J. and Sano, D. [1995], 'Sun web: user interface design for sun microsystem's internal web', *Computer Networks and ISDN Systems* **28**(1–2), 179–188. (Cited on page 93.)

[101] Nobel, P. A. and Shiffrin, R. M. [2001], 'Retrieval processes in recognition and cued recall', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **27**(2), 384–413. (Cited on page 138.)

[102] Pan, C. S., Shell, R. L. and Schleifer, L. M. [1994], 'Performance variability as an indicator of fatigue and boredom effects in a vdt data-entry task', *International Journal of Human-Computer Interaction* **6**(1). (Cited on page 78.)

[103] Pantano, V., Cardew-Hall, M. J. and Hodgson, P. D. [2002], Technology diffusion and organisation culture: Preliminary findings from the implementation of a knowledge management system, *in* T. S. Durrani, ed., 'Engineering Management Conference 2002 (IEMC'02)', IEEE International, pp. 166–171. (Cited on pages 14, 23, and 24.)

[104] Pejtersen, A. M. and Albrechtsen, H. [2000], Ecological work based classification schemes, *in* C. Beghtol, L. Howarth and N. J. Williamson, eds, 'Dynamism and Stability in Knowledge Organization: Proceedings of the Sixth International ISKO Conference', Vol. 7 of *Advances in Knowledge Organization*, Ergon-Verlag, Würzburg, Germany. (Cited on page 87.)

[105] Pejtersen, A. M. and Fidel, R. [1998], 'A framework for work centered evaluation and design: A case study of IR on the web'. Working paper for MIRA workshop, Grenoble, March 1998. Retrieved 23 May 2007.
URL: *http://www.dcs.gla.ac.uk/mira/workshops/grenoble/fp.pdf*
(Cited on page 88.)

[106] Porter, J. [2005], 'Read this post because it has zeldman in the title'. Retrieved 25 July 2006.
URL: *http://bokardo.com/archives/zeldman-in-the-title* (Cited on page 125.)

[107] Ranganathan, S. R. [1937], *Prolegomena to Library Classification*, 1st edn, Asia Publishing, New York. (Cited on page 89.)

[108] Ranganathan, S. R. [1967], *Prolegomena to Library Classification*, 3rd edn, Asia Publishing, New York. (Cited on page 89.)

[109] Raskin, J. [2000], *The Humane Interface: New Directions for Designing Interactive Systems*, Addison Wesley Professional. (Cited on page 64.)

[110] Rosch, E. H. [1973], 'Natural categories', *Cognitive Psychology* **4**, 328–350. (Cited on page 33.)

[111] Rosch, E. H. [1975], 'Cognitive representations of semantic categories', *Journal of Experimental Psychology: General* **104**(3), 192–233. (Cited on pages 33, 48, and 52.)

[112] Rosch, E. H. and Mervis, C. B. [1975], 'Family resemblances: Studies in the internal structure of categories', *Cognitive Psychology* **7**(4), 573–605. (Cited on pages 33 and 48.)

[113] Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. and Boyes-Braem, P. [1976], 'Basic objects in natural categories', *Cognitive Psychology* **8**(4), 382–439. (Cited on pages 31, 35, and 52.)

[114] Rowley, J. [1994], 'The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research', *Journal of Information Science* **20**(2), 108–119. (Cited on page 109.)

[115] Rugg, G. and McGeorge, P. [1997], 'The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts', *Expert Systems* **14**(2), 80–93. (Cited on page 93.)

[116] Schmitz, P. [2006], Inducing ontology from flickr tags, *in* 'Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland'. Retrieved 3 July 2006.
URL: *http://www.rawsugar.com/www2006/22.pdf* (Cited on page 150.)

[117] Sebastiani, F. [2002], 'Machine learning in automated text categorization', *ACM Computing Surveys* **34**(1), 1–47. (Cited on pages 95, 97, 98, 99, 102, 105, and 106.)

[118] Segal, R. B. and Kephart, J. O. [1999], MailCat: an intelligent assistant for organizing e-mail, *in* O. Etzioni, J. P. Müller and J. M. Bradshaw, eds, 'AGENTS '99: Proceedings of the third annual conference on Autonomous Agents', ACM Press, New York, NY, USA, pp. 276–282. (Cited on page 190.)

[119] Sherman, J. W., Lee, A. Y., Bessenoff, G. R. and Frost, L. A. [1998], 'Stereotype efficiency reconsidered: Encoding flexibility under cognitive load', *Journal of Personality and Social Psychology* **75**(3), 589–606. (Cited on page 70.)

[120] Shirky, C. [2006], 'Ontology is overrated: Categories, links, and tags.'. Retrieved 3 January 2007.
URL: *http://www.shirky.com/writings/ontology_overrated.html*
(Cited on pages 112 and 114.)

[121] Smith, E. E., Balzano, G. J. and Walker, J. [1978], Nominal, perceptual, and semantic codes in picture categorization, *in* J. W. Cotton and R. L. Klatzky, eds, 'Semantic Factors in Cognition', Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA. (Cited on page 31.)

[122] Smith, E. E. and Medin, D. L. [1981], *Categories and Concepts*, Cognitive Science Series, Harvard University Press, Cambridge, Massachusetts, USA. (Cited on page 32.)

[123] Smith, J. I., Cardew-Hall, M., Pantano, V. and Hodgson, P. D. [2004], Design, implementation and use of a knowledge acquisition tool for sheet metal forming, *in* '24th ASME Computers In Engineering Conference (DETC 2004), Salt Lake City, USA'. (Cited on pages 13, 14, 15, 25, and 26.)

[124] Smith, J. I., Pantano, V., Cardew-Hall, M. and Hodgson, P. [2003], A software development methodology for implementing knowledge-based software on the manufacturing shop floor, *in* J. Hodson, ed., '9th International Conference on Manufacturing Excellence (ICME 2003)'. (Cited on pages 14 and 15.)

[125] Solomon, P. [2000], Exploring structuration in knowledge organization: Implications for managing the tension between stability and dynamism, *in* C. Beghtol, L. Howarth and N. J. Williamson, eds, 'Dynamism and Stability in Knowledge Organization: Proceedings of the Sixth International ISKO Conference', Ergon, Würzburg, Austria, pp. 254–260. (Cited on pages 41 and 45.)

[126] Spiteri, L. [1998], 'A simplified model for facet analysis', *Canadian Journal of Information and Library Science* **23**(1–2), 1–30. (Cited on pages 89 and 90.)

[127] Star, S. L. [1989], The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving, *in* L. Gasser and M. N. Huhns, eds, 'Distributed Artificial Intelligence', Pitman, London, England, pp. 37–54. (Cited on page 72.)

[128] Star, S. L. and Strauss, A. [1999], 'Layers of silence, arenas of voice: The ecology of visible and invisible work', *Computer Supported Cooperative Work (CSCW)* **8**(1–2), 9–30. (Cited on page 78.)

[129] Steinbach, M., Karypis, G. and Kumar, V. [2000], A comparison of document clustering techniques, Technical Report #00-034, Department of Computer Science and Egineering, University of Minnesota. (Cited on pages 152, 154, 155, and 162.)

[130] Strehl, A. and Ghosh, J. [2000], A scalable approach to balanced, high-dimensional clustering of market-baskets, *in* M. Valero, V. Prasanna and S. Vajapeam, eds, 'Proceedings of the Seventh International Conference on High Performance Computing (HiPC 2000), 17-20 December 2000, Bangalore, India', Vol. 1970 of *Lecture Notes in Computer Science*, Springer, Berlin, pp. 525–536. (Cited on pages 147, 152, 153, and 154.)

[131] Suchman, L. [1994], 'Do categories have politics? the language/action perspective reconsidered', *Computer Supported Cooperative Work (CSCW)* **2**, 177–190. (Cited on page 74.)

[132] Surowiecki, J. [2005], *The Wisdom of Crowds*, Anchor, New York. (Cited on pages 48 and 67.)

[133] Svenonius, E. [1986], 'Unanswered questions in the design of controlled vocabularies', *Journal of the American Society for Information Science* **37**(5), 331–340. (Cited on pages 81, 108, and 115.)

[134] Tennis, J. T. [2005], 'Experientialist epistemology and classification theory: Embodied and dimensional classification', *Knowledge Organization* **32**(2), 79–92. (Cited on pages 85, 86, and 89.)

[135] Thomas, C. F. and Griffin, L. S. [1999], 'Who will create metadata for the internet?', *First Monday* **3**(12). (Cited on pages 48 and 111.)

[136] Tullis, T. and Wood, L. [2004], How many users are enough for a card-sorting study?, *in* D. Wilson and C. Z. Josephson, eds, 'Connecting Communities: Usability Professionals Association Conference 2004 (UPA 2004)'. Retreived 23 May 2007.
URL: *http://websort.net/articles/Tullis&Wood.pdf* (Cited on page 94.)

[137] Upchurch, L., Rugg, G. and Kitchenham, B. [2001], 'Using card sorts to elicit web page quality attributes', *IEEE Software* **18**(4), 84–89. (Cited on page 94.)

[138] van der Aalst, W. M. P. [1998], 'The application of petri nets to workflow managment', *Journal of Circuits, Systems and Computers* **8**(1), 21–66. (Cited on page 73.)

[139] Vander Wal, T. [2005], 'Folksonomy definition and wikipedia'. Retrieved 25 July 2006.
URL: *http://www.vanderwal.net/random/entrysel.php?blog=1750* (Cited on page 125.)

[140] Vander Wal, T. [2006], 'Memephoria'. Retrieved 27 June 2006.
URL: *http://www.vanderwal.net/random/entrysel.php?blog=1829* (Cited on page 125.)

[141] Voorbij, H. J. [1998], 'Title keywords and subject descriptors: a comparison of subject search entries of books in the humanities and social sciences', *Journal of Documentation* **54**(4), 466–476. (Cited on page 109.)

[142] Wang, K., Xu, C. and Liu, B. [1999], Clustering transactions using large items, *in* 'Proceedings of the 1999 ACM CIKM International Conference on Information and Knowledge Management', ACM Press. (Cited on page 162.)

[143] Weinberger, D. [2005], 'Taxonomies and tags: From trees to piles of leaves', *Release 1.0* **23**(2). (Cited on page 112.)

[144] Wittgenstein, L. [1953], *Philosohpical Investigations*, Basil Blackwell, Oxford. Translated by G. E. M. Anscombe. (Cited on page 86.)

[145] Wright, G. and Ayton, P. [1987], 'Eliciting and modelling expert knowledge', *Decision Support Systems* **3**(1), 13–26. (Cited on page 93.)

[146] Yang, Y., Guan, X. and You, J. [2002], Clope: A fast and effective clustering algorithm for transactional data, *in* 'KDD 2002: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM Press, pp. 682–687. (Cited on pages 152 and 163.)

[147] Zeldman, J. [2005*a*], 'Remove forebrain and serve'. Retrieved 23 June 2006.
URL: *http://www.zeldman.com/daily/0505a.shtml* (Cited on page 125.)

[148] Zeldman, J. [2005*b*], 'Tag clouds are the new mullets'. Retrieved 23 June 2006.
URL: *http://www.zeldman.com/daily/0405d.shtml* (Cited on page 124.)