

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

James Sinclair: *Categorisation for Small–Medium Information Systems—
An Exploration*, A thesis submitted for the degree of Doctor of Philos-
ophy of 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 your 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

1	INTRODUCTION	1
1.1	Thesis Overview	5
1.2	Original Contributions	6
I	THE CATEGORISATION PROBLEM	9
2	CASE STUDY: CATEGORISATION IN A MANUFACTURING ENVIRONMENT	11
2.1	Background	11
2.1.1	Sheet Metal Forming	12
2.1.2	The Knowledge Management System	14
2.1.3	Categorisation in SIMPRESS	16
2.2	The Study	18
2.3	Results from SIMPRESS	19
2.3.1	Use of the <i>Other</i> Category	19
2.3.2	Uneven Category Usage	21
2.3.3	System Design Issues	22
2.4	Results from Interviews	22
2.5	Discussion	26
2.6	Conclusion	27
3	COGNITIVE REASONS FOR CATEGORISATION PROBLEMS	29
3.1	Conceptions of Categorisation	30
3.1.1	Metaphorical Understanding and Categorisation	30
3.1.2	The Classical Theory of Categorisation	31
3.2	Category Structure in the Mind	33
3.2.1	Graded Structure	33
3.2.2	Fuzzy boundaries	34
3.2.3	Basic Level Categories	34
3.3	Classification versus Categorisation	35
3.3.1	Classification	35
3.3.2	Categorising and Classifying Information	37
3.4	Situated Learning and Category Dynamics	39
3.5	The Vocabulary Problem	42
3.6	Implications	44
3.6.1	Category Architecture	44
3.6.2	Category Dynamics	44
3.6.3	Vocabulary	45
3.7	Summary	45
4	GRADED CATEGORISATION AND USER INTERFACES	47
4.1	Introduction	47
4.1.1	Graded Categorisation	48
4.2	Method	49
4.2.1	Game Procedure	50
4.2.2	Category Selection	52
4.2.3	Selection of information artifacts	52
4.2.4	Participants	53

4.3	Results	54
4.3.1	Categorisation Results	54
4.3.2	Categorisation Accuracy	61
4.3.3	Time To Categorise	63
4.4	Discussion	66
4.4.1	Categorisation Accuracy	66
4.4.2	Limitations of the Study	67
4.5	Conclusion	68
5	CONTEXTUAL REASONS FOR CATEGORISATION PROBLEMS	69
5.1	Why Categorise?	70
5.1.1	Cognitive Categories	70
5.1.2	Concrete Categories	71
5.2	Causes of Categorisation Problems	73
5.2.1	Conflicting Requirements	73
5.2.2	Political & Social Consequences	74
5.2.3	Perceptions of the System	75
5.2.4	Interpretation and Subjectivity	76
5.2.5	Environmental Dynamics	77
5.2.6	Prediction	77
5.2.7	The Tedium of Data Entry	78
5.2.8	Summary	79
5.3	Implications for Information System Design	79
5.4	Problem Definition	82
6	POTENTIAL SOLUTIONS	85
6.1	Ecological Classification Schemes	85
6.2	Faceted Analysis	89
6.3	Card Sorting	93
6.4	Automatic Text Classification	95
6.4.1	Rule Based Systems	96
6.4.2	Pattern Matching Systems	97
6.4.3	Advantages and Disadvantages	106
6.5	Uncontrolled Vocabularies	107
6.5.1	Free Text Search	108
6.5.2	Author-Supplied Metadata	110
6.6	Folksonomies	111
6.6.1	Advantages of Folksonomies	112
6.6.2	Disadvantages of Folksonomies	115
6.7	Conclusion	116
II	FOLKSONOMIES	119
7	INVESTIGATION OF FOLKSONOMY TAG CLOUDS	123
7.1	Introduction	123
7.1.1	Tag Clouds	124
7.1.2	The Study	126
7.2	Method	126
7.2.1	Characteristics of the dataset	131
7.3	Results	133
7.3.1	Last-strike queries	133
7.3.2	Browsing versus finding	133

7.3.3	Presence of relevant keywords in the tag cloud	135
7.3.4	Participants' preference	137
7.3.5	Tag cloud as a visual summary	138
7.3.6	Tag cloud occlusion	139
7.4	Conclusion	140
8	COMPARISON OF FOLKSONOMY CLUSTERING TECHNIQUES	143
8.1	Introduction	143
8.2	Background	144
8.2.1	Why Cluster Folksonomies?	145
8.2.2	Clustering Techniques	145
8.3	Previous Work	148
8.4	Method	151
8.4.1	External Cluster Quality Measures	152
8.4.2	My Approach	156
8.4.3	Choice of Datasets	158
8.4.4	Implementation of Algorithms	160
8.5	Results	164
8.5.1	Intra-cluster Similarity	164
8.5.2	Category Scatter	164
8.5.3	Q Measure	167
8.5.4	Clustering Results	171
8.6	Discussion	171
8.7	Conclusion	172
9	DESIGN OF A FOLKSONOMY-BASED SYSTEM	173
9.1	Motivation	173
9.2	Related Work	174
9.3	Using SocRef	176
9.3.1	Browsing	176
9.3.2	Data Entry and Export	181
9.4	System Design	184
9.4.1	System Architecture	184
9.4.2	Entering Information in SocRef	185
9.4.3	Finding Information in SocRef	190
9.4.4	Sharing in SocRef	192
9.5	Implementation	193
9.6	Discussion	193
9.6.1	When is a Folksonomy Useful?	193
9.6.2	Bulk Uploading	194
9.6.3	Clustering Improvements	195
9.7	Conclusion	195
10	SUMMARY, CONCLUSION AND FUTURE WORK	197
10.1	Summary	197
10.2	Conclusion	202
10.2.1	Contribution to Knowledge	203
10.3	Future Work	205
III	APPENDIX	207
A	CONCERN TYPES IN SIMPRESS	209
B	POPCAT INFORMATION SHEET	211

B.1	Categorisation	211	
B.2	The Study	211	
B.3	How do I Play?	212	
C	SLASHDOT CATEGORY FREQUENCIES		213
D	DATASET TAG FREQUENCIES	215	
E	CLUSTERING RESULTS	219	
	BIBLIOGRAPHY	225	

LIST OF FIGURES

Figure 1	A line of heavy presses used in sheet metal forming.	12
Figure 2	Activities involved in the sheet metal forming product life cycle.	13
Figure 3	Knowledge feedback loop facilitated by SIMPRESS	15
Figure 4	Data entry screens for loading an FMI in SIMPRESS	17
Figure 5	Number of FMIs raised by month	24
Figure 6	Categorisation versus Classification	37
Figure 7	The user interface for the categorisation game.	51
Figure 8	The two categorisation interfaces.	51
Figure 9	Histograms for each interface	54
Figure 10	Correlation between GI and NGI	59
Figure 11	Variance for GI versus mean DoM	60
Figure 12	Alternative GI	65
Figure 13	The multiple perspectives involved in Cognitive Work Analysis (CWA)	88
Figure 14	Screen capture from wine.com.	91
Figure 15	Steps in training an automatic classifier	98
Figure 16	Example of the Rocchio method of automatic classification	100
Figure 17	Example of the k nearest neighbours (k-NN) method of automatic classification	101
Figure 18	An example of a decision tree classifier.	102
Figure 19	Example rule for an inductive learner	103
Figure 20	Example representation of a neural network.	104
Figure 21	Example to illustrate support vector machines (SVMs)	105
Figure 22	An example of a tag cloud.	125
Figure 23	The interface for tagging articles.	128
Figure 24	The search interface showing the search box and tag cloud.	129
Figure 25	Example of results from a tag cloud query.	130
Figure 26	The exit Survey.	131
Figure 27	Query method used to answer each question	134
Figure 28	Queries required to answer each question	136
Figure 29	Articles inaccessible from the tag cloud.	139
Figure 30	Percentage of articles not accessible from the tag cloud.	140
Figure 31	Observed distributions of co-occurring tags	149

Figure 32	An example of a cluster histogram.	163
Figure 33	Intra-cluster similarity versus number of clusters for the Slashdot dataset.	165
Figure 34	Intra-Cluster Similarity versus number of clusters for the RawSugar dataset	166
Figure 35	Category scatter for each clustering algorithm.	167
Figure 36	Category Scatter for the RawSugar dataset	168
Figure 37	Raw quality measure for each algorithm and the random baseline	169
Figure 38	Q-measure for the RawSugar dataset.	170
Figure 39	Example of a circular tag cloud	176
Figure 40	The login page for Socref	177
Figure 41	The 'What People are Reading' page.	178
Figure 42	The tag page.	179
Figure 43	Page showing a summary of information for a single resource.	180
Figure 44	User interface for entering tags.	181
Figure 45	The 'My References' page	182
Figure 46	Interface for uploading a single BibTeX entry	183
Figure 47	Options for exporting a group of resources	184
Figure 48	People, Tags and Resources form a tripartite network	185
Figure 49	Manual resource entry interface	189
Figure 50	The process of entering a resource into SocRef	190
Figure 51	The two categorisation interfaces.	212

LIST OF TABLES

Table 1	Indicative questions for the semi-structured interviews.	19
Table 2	Usage frequency for each concern type.	20
Table 3	FMIs labelled <i>other</i> .	21
Table 4	Comparison of Categorisation and Classification	38
Table 5	Artefact Media Types	53
Table 6	Participants	54
Table 7	Categorisation Results: Artefact 1	55
Table 8	Categorisation Results: Artefact 2	55
Table 9	Categorisation Results: Artefact 3	56
Table 10	Categorisation Results: Artefact 4	56
Table 11	Categorisation Results: Artefact 5	56
Table 12	Categorisation Results: Artefact 6	57

Table 13	Categorisation Results: Artefact 7	57
Table 14	Categorisation Results: Artefact 8	57
Table 15	Categorisation Results: Artefact 9	58
Table 16	Categorisation Results: Artefact 10	58
Table 17	Pair-wise correlation values within homogenous groups	61
Table 18	Correlation between $d_{i,j,k}$ and $\hat{d}_{i,j,k}$	62
Table 19	Agreement percentage within homogenous groups	63
Table 20	Mean time taken to categorise each artefact	64
Table 21	Mean time to categorise by first language (seconds)	65
Table 22	Example questions asked at each level of Cognitive Work Analysis (CWA).	88
Table 23	Facets for an online wine store	90
Table 24	Example rules for a rule based classifier	96
Table 25	Discriminating Features	99
Table 26	Feature Vectors	99
Table 27	Summary of advantages and disadvantages for approaches in the literature	117
Table 28	Participants	127
Table 29	Questions asked of participants to elicit information-seeking behaviour.	129
Table 30	Last-Strike Queries	133
Table 31	Mean queries to answer question where participants relied on a single interface	135
Table 32	Last-Strike Queries when relevant keywords were present in tag cloud.	137
Table 33	Participants' preferences from the exit survey.	137
Table 34	Reasons for choosing one interface over another.	138
Table 35	Notation for external cluster quality measures	153
Table 36	Top ten tags and categories from the Slashdot dataset.	159
Table 37	Top ten tags and categories from the RawSugar dataset.	159
Table 38	Different types of resource	186
Table 39	Field types for a resource in SocRef	187

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