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