

CATEGORISATION IN KNOWLEDGE CONTEXTS

A mid-term review paper prepared by James Sinclair

Department of Engineering
The Australian National University

February 2006

INTRODUCTION

Categorisation is something that we do naturally and unconsciously every day. We recognise one animal as a cat and another as a dog. We organise objects in the world around us in ways that reflect these categories. In our kitchens, we keep baking trays with other baking trays, saucepans with other saucepans and keep food separate from cleaning products. We categorise ideas, people, tasks and objects. Categorisation is fundamental to the way we think.

Yet when we categorise things in information systems, problems frequently occur. The folders in our personal computer systems are often disorganised and messy. Databases seem to be full of entries labelled 'other'. The entry we are trying to file just does not seem to fit nicely into any of the available categories. If we categorise so naturally in our brains, why is it that our information systems are so disorganised?

In industries where knowledge is essential to maintaining competitive advantage, a disorganised information system can severely hamper workers. What good is a knowledge base if I cannot find the information I need within it?

The research described in this paper explores why this disorganisation is so prevalent and what we can do to minimise it. Beginning with a motivating study, we examine some of the common categorisation problems that arise in information systems supporting knowledge management. We then move on to explore some of the cognitive reasons why these problems occur. Following that we explore problems arising from the context in which categorisation occurs, such as political and social causes. Based on this understanding of the different causes of categorisation problems, we then define our research problem. Finally, we propose avenues of research to address this.

MOTIVATION

The motivation for this topic came out of working with a knowledge management system (KMS) developed by the ANU for an Australian automotive manufacturer. The idea of the system was that when problems occurred during manufacturing an operator could record what happened along with the action taken to fix it. In this way, operators can solve the problem quickly if it occurs again, or avoid it entirely in future. The system was unique in that it used a visual mark-up system and integrated smoothly with digital cameras. Instead of writing lengthy textual descriptions, operators could simply draw a red circle on a photo of the part to indicate what was wrong.

Although the photograph-based user interface worked well, a problem seemed to be occurring in the system. It seemed that many of the entries were poorly classified, making them difficult to find later. This led to an initial investigation into the KMS to find out what was happening. To begin with, we examined 201 of the most recent entries. When analysed, more than half of the entries were classified as “Other” (see Figure 1).

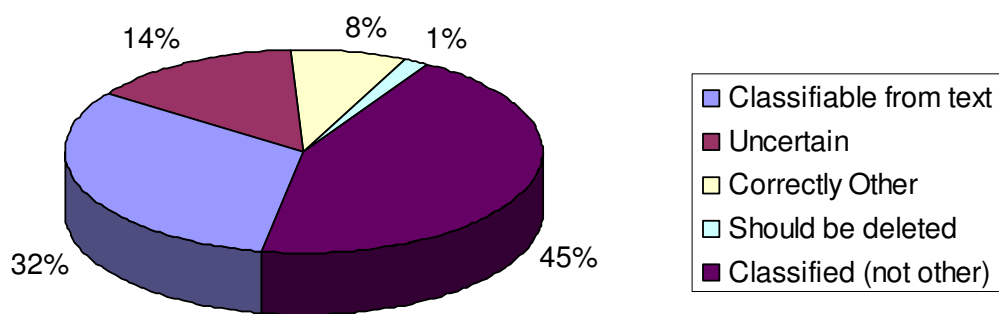


Figure 1. Classifications of KMS entries in the case study system.

Of the total, the researcher could have easily classified around a third of the entries using the existing classification scheme. Around 8% of the entries appeared to have been correctly classified “other”. This confirmed our suspicions that the classification system was not working as it should.

In order to find out what was happening, we conducted interviews with various employees who used the system. From this process, it emerged that two different kinds of problems seemed to be occurring:

1. The first kind of problem related to people's perceptions of the system. The interview process revealed that the engineers working in this area were very busy people, and consequently delegated things like data-entry tasks to sub-ordinates or work-experience students. Hence, the people making classifications were not the people with expert knowledge related to the entry. To make matters worse, some of the sub-ordinates had received inadequate training in using the system. Engineers viewed making entries into the KMS as administrative work, and 'non-essential' to the core task of making cars.
2. The second kind of problem related to the classification scheme itself. In many cases, the correct classification for an entry seemed to be ambiguous. The available classifications were broken down by problem-type, such as 'splits', 'wrinkles', 'burs', 'fouling', 'weld-integrity', 'spring-back', etc. There were also classes for problems related to 'design', or 'CAD', or 'process improvements'. Often a problem would seem to fit quite well into multiple categories, or sometimes *sort-of* fit into a few classes, but *not really* fit well into one only. For example, in some entries the operator had written "CAD/Design issue" in the text for the entry, yet an entry could only be either a "CAD" issue or a "design" issue. Should the operator classify it as "design" or as "CAD"? Or should the problem be classed "Other" if it does not fit neatly into either class?

At first, we assumed that the first kind of problem was simply a result of inadequate training and the unique environment of that particular manufacturing plant. Hence, the most promising avenue of investigation appeared to be the second type of problem. This led us to look at categorisation at a cognitive level to understand why classifying things in this manner was often difficult.

CATEGORISATION AT THE COGNITIVE LEVEL

Categorisation is fundamental to the way our minds work, and hence plays a significant part in nearly every aspect of our lives. Exactly how our minds categorise things, however, is still a matter of research and debate. For those not studying cognitive science (and even for some of those who are), the waters are even further muddied by confusion over the difference between categorisation and classification.

CATEGORISATION AND CLASSIFICATION

Much of the time the terms 'categorisation' and 'classification' are used interchangeably when they are, in fact, two distinct concepts. In a paper addressing this issue specifically, Jacob (2004) defines categorisation as follows:

Categorisation is 'the process of dividing the world into groups of entities whose members are in some way similar to each other.'

Classification, on the other hand, is the process of dividing a set of entities into mutually exclusive classes related according to formally defined rules. Jacob uses the following definition:

'A classification scheme is a set of mutually exclusive and nonoverlapping classes arranged within a hierarchical structure and reflecting a predetermined ordering of reality.'

These definitions give rise to some distinct properties of classification as opposed to categorisation:

- Classes are mutually exclusive. This means that an item can belong to one and only one class. Libraries are a good example of this, as a book can only be in one physical location on the shelf. It cannot sit on two different shelves at once.
- A predetermined set of principles assigns each item to a class. That is, there is a set of rules defining what does or does not belong in each category. This means that no one member of a class can be a better example of that class than any other member is.
- Hierarchical structure implies that all sub-classes must share the defining properties of the super-class.

Examples of classification schemes in use include biological taxonomies, the Dewey Decimal classification scheme, and the periodic table of elements.

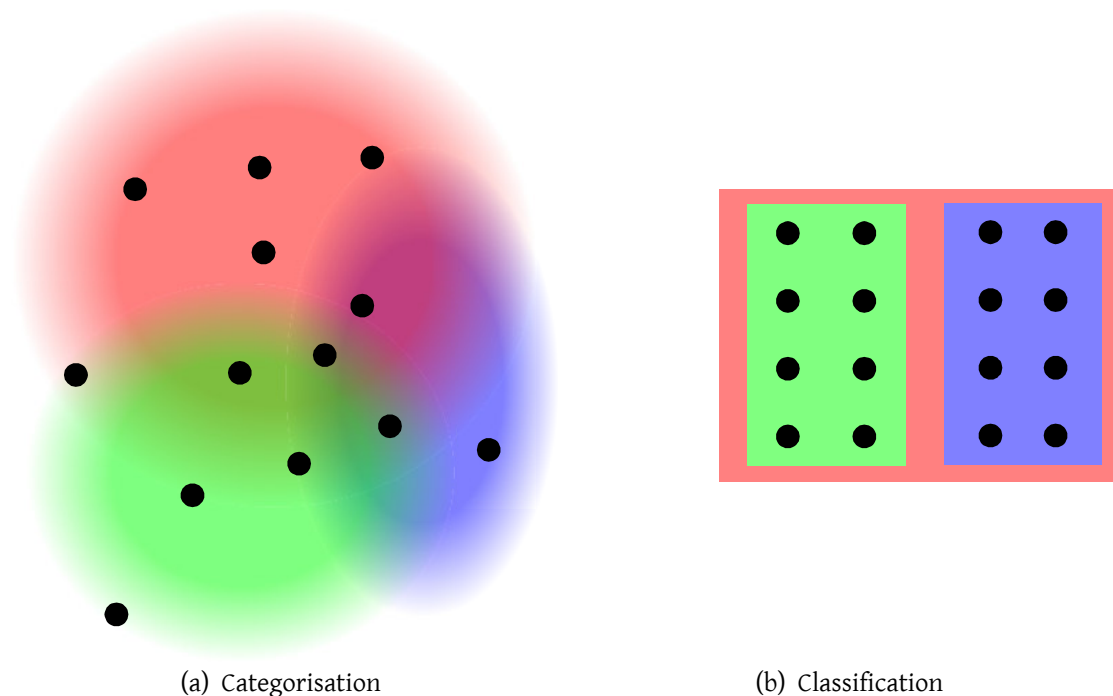


Figure 2. (a) Categorisation does not require rigid boundaries and some items may belong to a category more than others do, (b) Classification requires rigid boundaries between classes and is hierarchical.

Categorisation, on the other hand, is what we do at a cognitive level to make sense of our world (Lakoff 1990). It can be messy and seemingly inconsistent, whereas classification is ordered, hierarchical and seemingly logical (see Figure 2). Each individual will categorise according to his/her own experiences, culture, and world-view. Many of the problems in

classification and categorisation arise from because we often try to make categorisations fit into classification schemes. There is a mismatch between categorisation in our heads and the classification schemes we use to organise things.

In order to help illustrate this distinction Weinberger (2005) compares classification systems to trees. Each leaf belongs to a single branch, and each branch attaches to another branch right up to the trunk. Categorisation schemes, on the other hand, are more like piles of leaves.

Categorisation is much broader than classification, and in some ways, we can think of classification as a sub-set of categorisation. Hence, when talking about both classification and categorisation in general, we use the term categorisation to cover both.

What we do in our heads, is categorise. Most information systems however, require us to classify. When we try to map fuzzy, unstructured categories onto hierarchically ordered, rigid classes, then problems quite naturally arise. These observations led us to investigate the implications for user interfaces.

USER-INTERFACE INVESTIGATION

The classification/categorisation distinction suggests that if we can make a classification scheme more like categorisation, then it might reduce some of the difficulties in assigning items to classes. To investigate this idea, we conducted an experiment comparing two user interfaces for categorisation: one using slider bars to indicate membership, while the other used radio buttons (as shown in Figure 3).

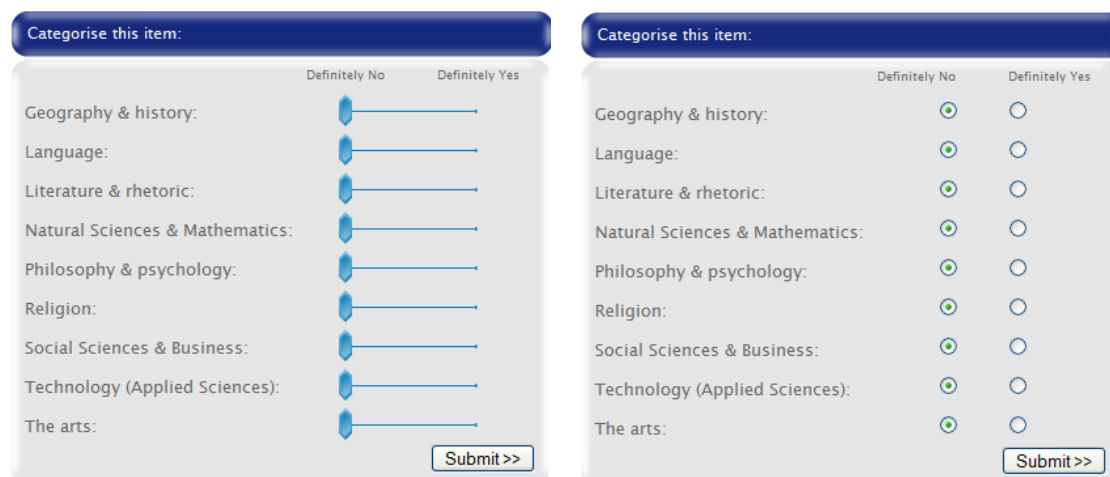


Figure 3. Categorisation interfaces compared in the user-interface study.

The slider-bar interface allowed users to indicate varying degrees of category membership, while the radio buttons only allowed users to indicate whether an item *did* or *did not* belong to a category. The hypothesis tested was that users who categorised using the slider-bar interface would show a greater degree of consensus than those who used the radio buttons.

To conduct the experiment, we asked participants to read/view a number of multimedia articles which included pictures, video and text items. After reading/viewing each article, participants assigned the article into categories using one of the two user interfaces shown

in Figure 3. In order to compare the two interfaces, the system recorded categorisation choices and time taken to read and categorise for each participant.

The results of this study confirmed the hypothesis that participants who used the slider-bar interface would show a greater degree of consensus in categorisation choices (see Figure 4). There was, however, a trade-off in terms of the time taken to categorise. Users of the slider-bar interface were on average over 8 seconds slower than users of the radio-button interface. Naturally, the slider bars require more mouse movements to operate, and hence would take longer, however, observations of participants during the experiment seemed to indicate that users of the slider-bar interface tended to deliberate more over their categorisation choices.

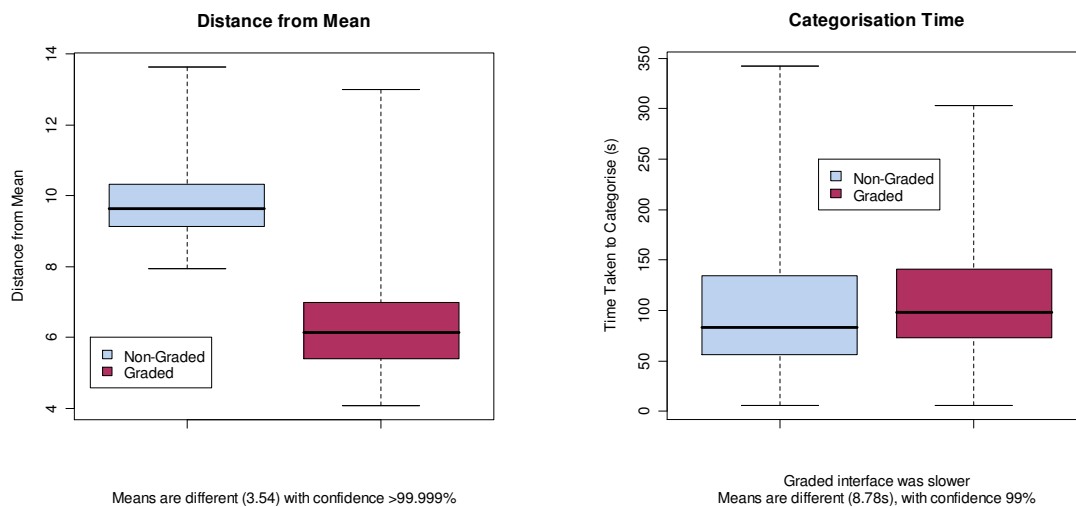


Figure 4. Results from the user interface study. The slider bar interface resulted in greater consensus, however users were slower using this interface.

These results are of interest because the trade-off between speed of data entry and categorisation accuracy is important for designers of information systems in business environments. The choice of interface used for categorisation needs to be aligned with business needs and goals. To what degree can the business afford to trade off accuracy against speed of data entry? Using slider bars may improve the accuracy of categorisation, but is it worth the extra time taken?

CATEGORISATION AT THE CONTEXTUAL AND STRUCTURAL LEVELS

While the user-interface investigation showed that an understanding of categorisation at the cognitive level has important implications for system design, the experiment itself did not immediately suggest any further avenues of study. In reading literature from a range of disciplines, however, it became apparent that the problems initially assumed to be unique

to the manufacturer in our case study were, in fact, common amongst a large number of organisations. The literature of Technology Adoption, Library and Information Science, and Sociology of Science all reported similar issues related to categorisation and the use of information systems. To understand these issues better, we first examine why we categorise.

PURPOSES FOR CATEGORISING AND CLASSIFYING

Before one can understand the problems associated with categorisation and classification, one must first understand why we categorise things in the first place. At a cognitive level we categorise in order to 'handle the variety and complexity of [...] day-to-day interactions with the environment' (Jacob 2004). At a higher level, however, we create categorisation and classification schemes for a number of different reasons:

- *To explicate procedures.* In order to specify what procedures to follow in certain situations, it is necessary to classify the situations themselves. For example, ANU Security distributes an emergency booklet that describes procedures to follow in case of fire, chemical spill, hostile situation, armed holdup, bomb threat, or other emergency. For each class of emergency, it gives different instructions on how to respond. These procedure classifications often take the form of 'when this happens, perform these actions', but levels of complexity can vary greatly, from simple things like the emergency booklet, to sophisticated decision trees with hundreds of classifications.
- *To gain understanding.* Another reason for categorising is simply to allow the organisation of thoughts on a topic in order to gain understanding (Kwasnik 1999). We create taxonomies and classifications in science for this purpose. For instance, in writing this paper we have classified different reasons for categorisation in order to gain an understanding of why people categorise. Zoologists and biologists arrange plants and animals into taxonomies in order to understand them better. Categorisation allows us to think and reason about groups of items and is hence an essential part of sense-making processes.
- *For numerical analysis and comparison.* Counting anything requires some form of categorisation. Sometimes this is as simple as counting apples, as opposed to oranges. Other times it can be a more complicated process, such as counting international causes of death (*cf.* Bowker and Star 1999).
- *To enable communication.* Oftentimes people use categorisation and classification schemes to negotiate a common vocabulary or nomenclature about a subject. An agreed-upon categorisation or naming scheme allows different groups of people to communicate about common objects. Ontologies and professional thesauri are examples of this kind of categorisation work.
- *To organise items and enable efficient retrieval.* This is where we categorise items in order to help us find them later. For example, many of us keep our pots and pans together in one spot in our kitchen. We keep knives separate from forks but group cutlery in general together. Why? So we can find them later. We do the same thing with computer files on our hard-drives. We create folders and sub-folders so we can find our documents when we need them again.

In this thesis, we focus on this last purpose: organising items to enable efficient retrieval.

CONTEXTUAL AND STRUCTURAL PROBLEMS IN CATEGORISATION

We now turn to the causes of categorisation problems that are not simply the result of cognitive mismatch. Many categorisation problems arise out of the complex environment in which categorisation occurs. These, we refer to as contextual causes. Other problems relate more to the design of the categorisation scheme and information system. These, we refer to as structural causes.

Contextual Causes

Contextual causes refer to problems arising from the organisational, social, and political environment in which information systems are used. Examples of contextual causes include:

- *Continual change within organisations.* As an organisation changes, the nature of what is kept in the information system will change with it. People leave, others arrive; product lines change; processes change. Thus, the categories that were useful last year may quickly become obsolete this year.
- *Attitudes of stakeholders.* If there is a perception that the data collection is unimportant or extraneous to a worker's central task, then quality of categorisation will suffer. Furthermore, the perceived value that management places on the data entry will also have a significant impact. If such work is unrewarded and unrecognised, then there is little motivation to perform the task with great care or precision.
- *The tedium of data entry.* Data entry is often boring and tiresome. Unfortunately, bored, tired workers tend to make more mistakes and cut corners where they can, and inevitably the quality of categorisation suffers.
- *Conflicting needs of stakeholders.* Often those entering the data are not the same ones using or analysing it. Different groups may use a classification scheme for a variety of purposes—all of which have conflicting requirements.
- *Political and social consequences of categorisation.* Often categorising in a particular way will have consequences for those performing the categorisation (Bowker and Star 1999). A common example is research grant applications. Often researchers will attempt to categorise their work as belonging to more popular areas of research in order to attract more funding, while in other areas they may categorise their work completely differently.

Structural Causes

Structural causes refer to problems arising from the design of the information system and categorisation scheme. Of course, the design of the information system and categorisation scheme forms part of the context in which categorisation occurs, hence they are also contextual causes. We distinguish structural causes however, because these are issues over which the system designer has a greater degree of control. While it is essential for a system designer to be aware of contextual issues, there is often little that the designers themselves

can do to change them. Structural issues, however, are an area where the system designer is able to have an impact. Examples of infrastructure causes include:

- *Poor user interface design.* As mentioned above, data entry is often boring and tiresome. A poorly designed user interface can cause frustration and increase the difficulty of an already tedious task.
- *Differences in vocabulary or 'world view.'* Very rarely are system designers the end users of the product they are designing. They will hence not have the same domain knowledge or as complete an understanding of the 'way things work' as the end users. Where there is a significant mismatch, the end-users will create work-arounds or modify the system to suit their particular understanding of the system.
- *Poor choice of granularity for data collection.* Too fine a detail results in an enormous set of categories that is unwieldy and difficult to use. Too coarse a detail means that important data may be lost.
- *Prediction of future needs.* Often the ideal categorisation scheme does not become apparent until the information system has been implemented and categorisation begun. Designing a categorisation scheme always involves prediction of what will be important. This also implies a decision as to what is not important to record.
- *Lack of system adaptability.* As described above, organizations continually change. A categorisation scheme that cannot be adapted to respond to changing needs will cause problems very quickly.

PROBLEM DEFINITION

Recall that the motivation for this study arose from problems observed in a small-scale KMS that incorporated large amounts of image data. In light of this and the categorisation issues observed at the cognitive, contextual and structural levels, we develop the following problem definition for the thesis.

The aim is to develop means and methods for categorising in smaller scale KM systems that meet the following requirements:

- *Does not require a dedicated librarian/administrator* to maintain and modify the classification/categorisation structure.
- *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 KMS operates changes over time. It must also be able to respond to corresponding changes in vocabulary.
- *Suitable for multi-media data.* Given that the KMS described relied heavily on digital photographs and scanned blueprints (among other things), the categorisation method must be suitable for non-textual data.

- *Suitable for small numbers of records.* The system should be able to work effectively with as few as 200 records.

APPROACHES IN THE LITERATURE

Current approaches to categorisation issues such as those described above come from a variety of academic backgrounds. Library and Information Science practitioners advocate Facet techniques and domain analysis. Human-Computer Interaction experts and Information Architects recommend the use of Card-Sorting to uncover users' mental models. Computer Scientists have contributed algorithms for automatic categorisation, which enable large bodies of documents to be classified without human intervention. This section gives an overview of these various approaches.

DOMAIN ANALYSIS AND BOUNDARY OBJECTS

In a review paper on classification literature, Mai (2004) identifies two trends in information science: '1) shifting from focusing on the systems and techniques, to the contexts and domains in which classifications function, and 2) shifting towards relativistic philosophies.' The current standard of practice is to conduct in-depth studies into the particular organisation, its people and its activities:

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. One needs to learn the language used in the community, since the classification must reflect and respond to this particular discourse community. A classification is not something that can be created *for* an organization by an epistemic authority; a classification must *grow out* of the organization. The classification is a typification of the language in the organization. (Mai 2004, emphasis in original)

This kind of approach is based on work by Hjørland and Albrechtsen (1995) in which they introduce Domain Analysis as a framework to approach information science. In short, the ideal way to construct a classification scheme, according to this view, is to conduct an ethnographic study of the organisation and base the classification scheme on the result.

Many of these ideas arose out of the work of sociologists such as Bowker and Star, who developed the idea of *boundary objects*. Boundary objects are entities that people who hold very different viewpoints have in common. The idea is that since the two different world-views overlap at this point, there can be a discourse based around this common object. Star compares the idea to a blackboard, which "sits in the middle" of a group of actors with divergent viewpoints.' (Star (1989) quoted in Albrechtsen and Jacob (1998)).

Classification schemes were identified as common boundary objects by Albrechtsen and Jacob (1998) and Bowker and Star (1999). As such, they recommend that the creation of a classification scheme should serve as a 'discursive arena', where the classification scheme

emerges as the result of discussion and debate amongst all stakeholders and accommodates many different points of view.

If followed, these approaches should result in an effective classification scheme that reflects the majority users' perspectives on the information. Coming primarily out of the library and information science literature, it is understandable however, that these approaches assume that a dedicated specialist is available to do the work of creating the classification scheme and performing classifications. In the kind of smaller-scale systems examined in this thesis, there is rarely such a trained professional available to devote their time to creating and maintaining the classification scheme.

FACETED ANALYTICO-SYNTHETIC THEORY

Faceted Analytico-Synthetic Theory (FAST) is often put forward as a solution to the problem of providing flexible, adaptable classification schemes.

The idea of Facet Analysis was first proposed by S. R. Raganathan in his book *Prolegomena to Classification* in 1933 and was further expanded and revised by the Classification Research Group (CRG) (Spiteri 1998). The approach gained significant popularity in Library and Information Sciences, and more recently has been applied to website design and information architecture (La Barre 2004b; La Barre 2004a).

The formal rules and principles for facet analysis, published by Raganathan and the CRG, are somewhat complicated and difficult to read (Spiteri 1998). This has led a number of authors to write simplified or summarized versions of the analysis process (Spiteri 1998; Kwasnik 1999; Denton 2003), each with their own modifications and variations. Thus, attempting to find a good introduction to the field can be a difficult process (La Barre 2004a).

Denton (2003) 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 basic premise is that 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.

It is claimed that faceted classification systems are flexible and adaptable. This stems from the requirement that facets be orthogonal and mutually exclusive. This means 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.

Once again, coming from the Library and Information Science 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.

AUTOMATIC CATEGORISATION

Automatic text categorisation (or, more correctly, classification) attempts to classify documents based on the textual contents of the document itself. Setting up an automatic classifier usually happens in a number of stages (shown in Figure 5).

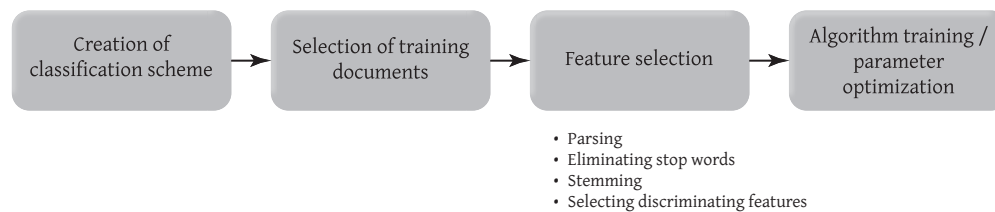


Figure 5. Steps involved in setting up an automatic classification system.

In the first stage, a human analyst creates a classification scheme for the set of documents to be categorised. If applying the technique to an already-existing collection, then the extant classification scheme can be used.

The next stage involves selecting representative documents from each class to train the classifier. When the automatic classifier operates, it decides which class a document belongs to based entirely on its similarity to the training set documents.

In the feature selection stage, the training documents are parsed in an attempt to extract the most salient terms or phrases on which to base the classifications. Usually this involves removing common words (called stop-words) and picking out words or phrases that help differentiate documents in a category from other documents.

The final phase involves tuning or training the classifier. How this works depends on the kind of classifier being used. For a Bayesian Modelling approach, this may involve performing a statistical regression analysis to determine which training document features produce the best classification. For a K-Nearest neighbour approach, it will involve optimising the cluster radius and minimum number of documents in a cluster.

There are a number of different classifiers that can be used in an automatic categorisation system. Lubbes (2003) gives an overview of some of these, such as K-nearest neighbour, Bayesian modelling, neural networks, support vector machines (SVMs), and rule-based approaches.

Automatic categorisation can be a very useful tool for dealing with large amounts of textual data. For the purposes of this thesis, however, it fails to address most of the major requirements. Automatic categorisation is normally applied after the classification scheme has already been created, and does not usually deal well with multimedia data. There is some potential, however, in examining clustering techniques developed in this field. This will be explored further later.

CARD SORTING

Card sorting is a relatively simple process often used as an aid to organising web-sites or in creating user-centred classification schemes. The idea is to write down a set of words or

phrases on a stack of index cards. This stack of cards is then given to a potential user or group of users, and they are asked to sort the cards into piles. Assuming they understand the concepts written on the cards, most people have no trouble arranging the cards into logical groups. In this way, the card sorting technique allows a system designer to 'discover users' mental model[s] of an information space' (Nielsen and Sano 1995).

The technique is particularly effective when people sort cards in a group. The group environment forces people to talk aloud about what they are doing and justify to others why a card might belong to a certain category. This can give an observer valuable insight into user perspectives on the information space.

Card sorting is a useful tool for uncovering users' perspectives on how information should be organised. Hence, it is a useful tool in creating a classification scheme. However, it does not guarantee a flexible, adaptable system that can operate without a dedicated administrator.

FOLKSONOMIES

Folksonomies are an unsupervised method of organising information that revolves around the concept of *tagging*. The users of a system such as Del.icio.us are encouraged to tag items (in this case, web pages) by assigning one or more keywords for their own personal use. For example, a user may tag a website selling environmentally responsible shoes with keywords such as 'shoes', 'shopping', 'ecofriendly' and 'fashion'. If the user wants to find that page again, they can do so using those tags. However, the tags for that web page are also made public so that any other user interested in ecofriendly shoes can also discover the page. In addition, any other users may add their own tags if they wish to remember the site for themselves. When many users do this, what emerges in the aggregate is 'a bottom-up, self-organized system for classifying [*sic.*] mountains of digital material' (Pink 2005).

The folksonomy concept has become quite popular in recent years. In 2005, the popularity of websites such as Del.icio.us¹ and Flickr² was enough to warrant articles on folksonomies in newspapers such as *The Guardian* (Burkeman 2005) and *The New York Times* (Pink 2005). Large companies such as Google and Yahoo! have bought folksonomy-based websites such as Flickr and Del.icio.us for large sums of money. The concept of grass-roots classification has captured the imagination of many.

Part of the attraction of folksonomies is that they are inherently user-centred. A professional with a library science degree does not perform the categorisation; instead, the users of the system categorise items in a manner that makes sense to them. In this way tagging is much more like categorisation than classification (Mathes 2004). The tags given to items are contextually relevant to the individual users, and may or may not be of relevance to others. For example, one of the top 100 tags on Del.icio.us is the tag 'toread', presumably assigned to web pages that users intend to read later. This particular tag is (mostly) only of interest to the individuals who apply it. Most tags, however, are descriptive and reflect a consensus as to what an item is about.

¹ <http://del.icio.us/>

² <http://www.flickr.com/>

Another distinguishing feature of folksonomies is their social aspect, as tags are made public to all users. A user tagging a web page with 'shoes' has a social incentive to tag the page so that others interested in shoes are able to find it too. In Del.icio.us, when a user tags a page, they are shown any tags other people have used for that page. This creates a social feedback mechanism where users are able to see what tags others use and adjust their own tags to match (they are also free not to do so).

Folksonomies have the most potential to address the requirements put forward in the problem statement described earlier. Since there is little overhead associated with a user creating new categories, folksonomies are able to evolve as the content and vocabulary of users changes with time. And since the categorisations are being performed by the users of the information, there is no need for a dedicated administrator to perform cataloguing. They are also well suited to multimedia data, as clearly demonstrated by the very popular Flickr website, which allows users to tag digital photographs.

Folksonomies do have a number of drawbacks however. First, there are all the problems associated with uncontrolled vocabularies, such as synonyms and ambiguity. One person may tag something 'nuts' referring to the edible variety, while another uses the same tag to refer to things that go with bolts; or, one person may use the tag 'apple' to describe their computer, while another person uses the tag 'macintosh' to refer to the same thing. This can make it difficult for users to attempt to retrieve information.

Another issue with folksonomies is that they generally rely on large numbers of people in order to work well. If there are not enough users tagging items, then there will tend to be little overlap in different tags people use. This is not necessarily a problem for the individual users themselves, since the tags they use are still useful to them. Some of the sharing and discovery benefits may be lost however and a shared vocabulary is not likely to emerge.

In spite of the disadvantages, folksonomies seem a good match for smaller scale KM systems. They do not require a dedicated administrator; they can evolve quickly as the organisation changes; and they tend to promote the sharing of information amongst users. The research being undertaken for this PhD, will focus on the usefulness of folksonomies for small scale KM systems.

CURRENT AND FURTHER RESEARCH

While folksonomies seem to have the potential to address the categorisation problem posed earlier, there are disadvantages associated with their use. The research described here aims to evaluate the use of folksonomies to support knowledge sharing, and to look for ways to minimise the disadvantages of folksonomies. Three areas of study are proposed below: 1) a case-study system, 2) user-interfaces for folksonomies, and 3) clustering folksonomy data.

FOLKSONOMY CASE STUDY

At the time of writing, a case study is underway investigating the use of folksonomies in KM systems. The case study system is an online database that allows academics and postgraduate students to post abstracts of papers they have read in order to keep track of citations and to share with others what they have been reading. Users can assign tags to papers and view tags that other people have assigned to papers.

The study aims to determine if folksonomies are effective when used in a small-scale system where users have widely varying interests. In such an environment, do people still have a tendency to use tags that others have used previously? Does the system encourage users to read articles they would not have discovered otherwise? If there is little in common between different users of the system, is it still of use even without the social aspects?

The study will involve a mixture of qualitative analysis of data recorded by the system, combined with semi-structured interviews with users of the system themselves.

FOLKSONOMY USER INTERFACE STUDY

A proposed study to be undertaken this year will look at user interface aspects of folksonomies. One common user interface element often associated with folksonomies is the tag cloud. A tag cloud lists the most popular tags in alphabetical order, and scales the font-size of each tag relative to its popularity (see Figure 6, for example). Each tag listed in the tag cloud is a hyperlink to a list of items tagged with that keyword.

All time most popular tags

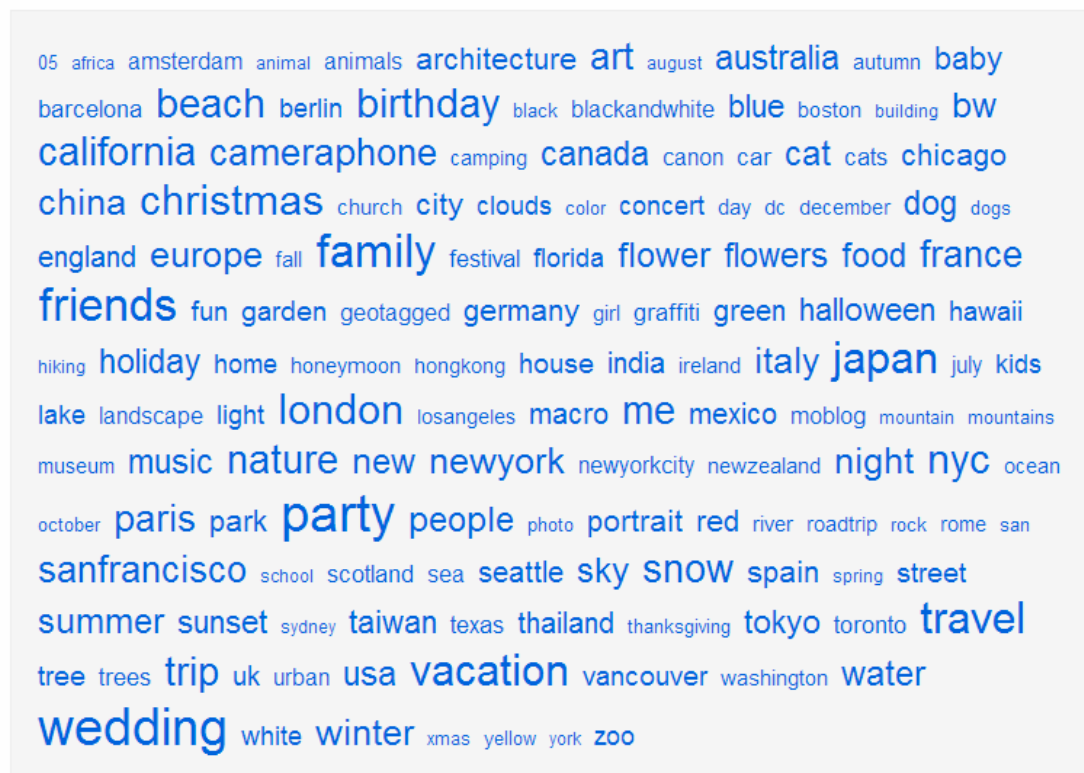


Figure 6. A tag cloud from Flickr.

While tag clouds seem to be popular, there is some debate over their usefulness as a navigation tool. Are they simply a pretty but useless toy, or do they present useful information to the user? Is the claim true that tag clouds may be of use for serendipitous searching but lack any usefulness in locating specific documents?

The user interface study will have two phases. The first phase will involve participants tagging news articles. This will create a folksonomy data set for the second phase. In the second phase, we will ask participants to carry out a number of information seeking activities using the folksonomy data set. In order to carry out these tasks they will be presented with the option of using a traditional keyword search interface, or a tag cloud. This will help us determine if the tag cloud is of any use in carrying out information seeking activities.

FOLKSONOMY CLUSTERING STUDY

As mentioned above, some of the problems associated with folksonomies are synonyms and ambiguity. However, it is possible that the use of clustering techniques could help ameliorate these problems. One group of items tagged 'apple' might also have the tags 'fruit', 'orchard' or 'granny smith'. Another item tagged 'apple' might also have the tag 'mac', 'OSX' or 'iBook'. Using a clustering algorithm could provide a user searching for 'apple' with different options based on these associated keywords.

Exactly how to apply clustering algorithms to folksonomy data remains to be explored. Which of the numerous clustering algorithms would work best? Which distance measure would provide the most useful grouping of tags? How can we present the results of clustering in a way that is intuitive and useful for users of the folksonomy?

The clustering study will investigate these questions using data gathered from previous studies and publicly accessible folksonomies. We will apply different algorithms with varying parameters to these data sets and evaluate their usefulness. We can then implement a clustering interface in the case study system mentioned above to investigate how users interact with automatically clustered groups.

CONCLUSION

Categorisation is a broad and complex area of study that crosses many disciplines. There are issues that arise from the complexity of the human brain, and other issues that arise from the context in which categorisation occurs. While there is an extensive body of research examining this issue, most of it assumes an expert administrator is available to keep things organised. This thesis examines the problem of how best to organise information in smaller scale knowledge management systems where such an administrator is not available.

The technique that shows most potential for addressing this issue is that of using folksonomies. Folksonomies have the double advantage of reflecting users' vocabularies,

and being able to evolve as organisations change. They are able to do this without requiring an administrator to manage categories.

There are some disadvantages to folksonomies however, and they are certainly not useful in all situations where categorisation is required. The proposed research seeks to investigate the usefulness of folksonomies when applied to smaller scale knowledge management systems. As part of this, we seek to determine how some of the disadvantages associated with folksonomies can be minimised.

REFERENCES

- Albrechtsen, H. and E. K. Jacob (1998). "The dynamics of classification systems as boundary objects for cooperation in the electronic library." *Library Trends* **47**(2).
- Bowker, G. C. and S. L. Star (1999). *Sorting Things Out: Classification and Its Consequences (Inside Technology)*, The MIT Press.
- Burkeman, O. (2005). Folksonomy. *The Guardian*, 12th September 2005.
- Denton, W. (2003). *How to make a Faceted Classification and Put it On the web*. <http://www.miskatonic.org/library/facet-web-howto.html>. (Accessed 7th March 2006). Last updated 13th August 2005 (Accessed 7th November 2006).
- Hjørland, B. and H. Albrechtsen (1995). "Toward a new horizon in information science: Domain-analysis." *Journal of the American Society for Information Science* **46**(6): 400-425.
- Jacob, E. K. (2004). "Classification and Categorization: A Difference that Makes a Difference." *Library Trends* **52**(3).
- Kwasnik, B. H. (1999). "The role of classification in knowledge representation and discovery." *Library Trends* **48**(1): 22.
- La Barre, K. (2004a). *Adventures in faceted classification: A brave new world or a world of confusion?* 8th International ISKO Conference: Knowledge organization and the global information society, London, 13-16 July 2004.
- La Barre, K. (2004b). *The Use of Faceted Analytico-Synthetic Theory as Revealed in the Practice of Website Construction and Design*. PhD Dissertation Proposal. Indiana University, Bloomington, USA. http://ella.slis.indiana.edu/~klabarre/ProposalLa_Barre.pdf.
- Lakoff, G. (1990). *Women, Fire, and Dangerous Things*, University Of Chicago Press.
- Lubbes, R. K. (2003). "So you want to implement automatic categorization?" *Information Management Journal* **37**(2): 60.
- Mai, J.-E. (2004). "Classification in context: relativity, reality, and representation." *Knowledge Organization* **31**.
- Mathes, A. (2004). *Folksonomies - Cooperative Classification and Communication Through Shared Metadata*. <http://adammathes.com/academic/computer-mediated->

[communication/folksonomies.html](#). Graduate School of Library and Information Science, University of Illinois Urbana-Champaign. Last updated December 2004 (Accessed 10th March 2006).

Nielsen, J. and D. Sano (1995). "Sun Web: user interface design for Sun Microsystem's internal web." *Computer Networks and ISDN Systems* **28**(1-2).

Pink, D. H. (2005). Folksonomy. *The New York Times*, 11th December 2005.

Spiteri, L. (1998). "A simplified model for facet analysis: Ranganathan 101." *Canadian Journal of Information & Library Science* **23**(1-2): 1-30.

Star, S. L. (1989). The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving. *Distributed Artificial Intelligence*. I. Gasser and M. N. Huhns. London, Pitman: 37-54.

Weinberger, D. (2005). "Taxonomies and Tags: From Trees to Piles of Leaves." *Release 1.0* **23**(2): 1-33.