A COMPARISON OF ALTERNATIVE APPROACHES FOR THE AUTOMATED ORGANISATION OF DESIGN INFORMATION

A. Lowe, C. A. McMahon and S. J. Culley

Abstract

Due to the ever-increasing amount of information available within modern engineering organisations, improved approaches need to be developed to allow for the automated organisation of documents (since pre-organised information can be more easily retrieved at search time). This paper presents a comparison of different approaches that can be used to organise textual design information within engineering organisations. Two types of information organisation strategies are presented and their advantages and disadvantages discussed: (i) *document clustering* and (ii) *document classification (or categorisation).*

Keywords: design information management, knowledge management, classification and retrieval

1. Introduction

Whilst computer technologies have made the creation of documents easier, and provided the means to open up huge digital collections to searching, in some ways this has paradoxically made the location of relevant information more difficult. There has been an explosive growth in the quantity of electronic documents that can be found within modern engineering organisations, and especially via the World Wide Web [1]. Increasingly the selection and screening of relevant and irrelevant information is critical and it is vital in developing ways of separating transitory information from important intellectual assets. The systematic organisation of textual design information into meaningful categories is a means to allow this to be achieved [2]. Simply put, by organising information *a priori* it can be more easily found. As noted by Foskett, "…(classifying documents) takes time; searching takes time. By increasing our effort at the classification stage – the input – we may well be able to reduce the amount of time we have to spend at the output stage in searching" [3].

Traditionally the organisation of information, to facilitate sharing and retrieval, has been associated with libraries and filing systems. With the development of computer-based information support systems, Web sites and company networks it has assumed increasing importance, but the sheer volume of electronic information is such that traditional manual approaches to classification are no longer able to cope. For this reason, approaches that permit the largely *automatic organisation* of digital documents need to be applied [4]. A discussion of these automated approaches and their application in design is the main focus of this paper.

2. Objectives

This paper presents a comparison of different approaches that can be used to organise textual design information within engineering organisations. The content of the paper is based on work carried out by the authors over the last few years, developing applications for organising text-based information within engineering companies. The purpose is not to provide a detailed

analysis of the various algorithms used, instead the aim is to give an overview of how these different approaches work in addition to a discussion of their various strengths and weaknesses. By describing the characteristics of various approaches, it should be possible for the reader to judge whether their use might be appropriate for a particular application.

3. Background

Any system for organising textual information comprises two core components:

- An ordering scheme consisting of groups or categories against which documents can be assigned.
- Some means of deciding how to assign documents with the groups or categories in the ordering scheme.

These two components can either be manually specified, or alternatively automatically inferred in some manner. In a manual classification process, people carry out the specification of both of these components. In automated information organisation processes, the degree to which the specification of these components can be carried out automatically is dependent on the selected strategy. In particular, a key distinction can be made between automated *document classification* and *document clustering* strategies:

- **Document clustering** – The grouping together of similar documents according to their textual content. Documents are organised into arbitrary groups – i.e. clusters – which are *not pre-defined*.
- **Document classification** (also termed *document categorisation*) – The assigning of documents to categories in a *pre-defined classification scheme or taxonomy*, according to the intellectual content of documents. The pre-defined scheme / taxonomy provides a model (of a domain) that identifies the important concepts / categories and structures them in a 'meaningful' way (often in the form of a classification hierarchy).

In automatic document classification the identification of the classification scheme is still carried out manually – i.e. the classification scheme is *pre-defined*. The differences between different classification approaches lie in the mechanisms and the algorithms used to decide how to assign documents with categories in the scheme. Document clustering is more demanding since there are no pre-existing categories (created by human experts) with which documents can be associated.

For this reason the term *unsupervised* is used to describe clustering techniques, in contrast to *supervised* classification techniques. Clustering can be viewed as having a greater number of degrees of freedom than classification – i.e. it is necessary to identify the clusters in addition to assigning documents with the clusters.

The schematic diagram in Figure 1, adapted from [5] and based on the authors' recent experiences, summarises some of the various performance characteristics of clustering and classification approaches. There can be seen to be a performance trade-off with the degree of automation and the accuracy of the information organisation approaches (i.e. how well documents are organised).

In the following sections of the paper more detailed discussions of various clustering (section 4) and automatic classification approaches (section 5) are given. This is followed by a discussion of the factors that affect the choice of information organisation strategies within engineering organisations and finally some key conclusions.

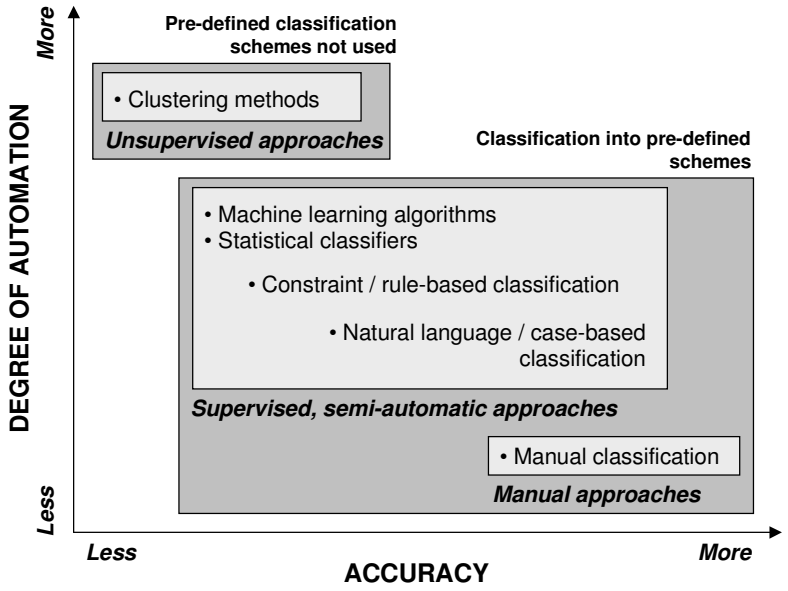

Figure 1 – Characteristics of different information organisation approaches

4. Document clustering

Clustering is a generic approach that allows the (automatic) grouping of similar objects. Furthermore, not only must similar objects be grouped, dissimilar objects must be kept distinct from each other. In an information retrieval context, the cluster hypothesis is that closely associated documents will tend to be relevant to the same requests. Thus, if one document is known to be of interest, then the whole cluster will also probably be of interest. Clustering can be viewed as an exercise in data reduction, whereby a whole set of individual objects (e.g. documents) can be represented by a smaller number of clusters representing those objects.

4.1 Document clustering methods

As previously noted, clustering methods do not classify documents into pre-defined categories. Instead, document clustering allows for the fully automatic *organisation* of documents into arbitrary groups, whereby those within a particular cluster are related by some metric. These similarity measures could include:

- The co-occurrence of similar terms or text strings inside documents.
- The frequency of co-occurrence of pairs of terms in documents.
- The co-occurrence of citations or hyperlinks.
- Various distance measures that assess the similarity of documents represented as vectors.

Details of clustering approaches are provided in the Information Retrieval literature [4][6][7]. There are basically two types of document clustering algorithms:

Partitioning clustering algorithms – These approaches divide collections of documents into a given number of smaller mutually exclusive sets of clusters. Typically, algorithms allow for the customisation of the number of clusters and a minimum and maximum size for each cluster.

Hierarchical clustering algorithms – These approaches provide sets of overlapping clusters at different 'resolutions' (as indicated in Figure 2). For computational reasons, agglomerative algorithms are generally favoured. These work as follows: *Step (i)* – each document is placed into its own cluster, *Step (ii)* – two clusters are chosen (via a 'linkage' criterion[1]) and merged and *Step (iii)* – if only one cluster is left then the process is complete, else Step (ii) is repeated.

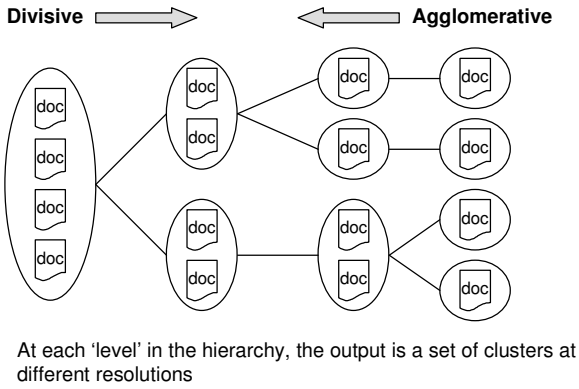

Figure 2 – Hierarchical clustering approaches

Note that variations of these basic types of algorithms have been developed that permit a degree of overlap between clusters. These overlapping clusters can be constrained to limit the number of objects that belong simultaneously to two clusters, or they can be unconstrained, allowing any degree of overlap in cluster membership.

The major advantage of these methods, as a means of organising documents, is that the process is carried out fully automatically, without the need for having previously identified a suitable classification scheme. However, whilst being fully automatic, clustering can be difficult to control and influence. Cluster results can tend to be rather arbitrary since the final clusters depend on the order in which documents are processed, the random selection of documents as initial cluster centres (if applicable) or the exact parameter values used. Rasmussen [7] also identifies the following problems with applying various clustering techniques in the text analysis domain:

- Difficulty in assessing the validity of the results (i.e. the clusters) obtained.
- Selecting appropriate attributes for clustering.
- Selecting an appropriate clustering method.
- High cost in terms of computational resources.

4.2 An example – clustering engineering papers

Table 1 provides an example of the results of a document clustering exercise carried out by the authors using the 390 papers from the proceedings of ICED'99. The documents were clustered with a partitioning clustering algorithm, using a commercial application supplied by a leading vendor of information management software[2]. The proprietary algorithm initially

[1] Note that the particular linkage criterion determines the characteristics of the clusters – common examples include 'nearest-neighbour', 'furthest-neighbour' and centroid linkage.

[2] The Search97 information management suite, provided by Verity, was used (www.verity.com).

suggested 6 clusters and also identified the keywords that were used to characterise the members of the clusters (shown in the second column).

Table 1: Clusters identified when organising ICED99 papers

	Keywords characterising the cluster
Cluster 1	Object, users, CAD, menus, gloves, scene, navigation, hand, 3-D, device
Cluster 2	Design, designer, functions, safety, system, approach, innovation, structure, environmental, risk
Cluster 3	model, products, design, engineering, simulation, mechanical, performance, implementation, variables, CAD
Cluster 4	method, development, companies, introduction, processes, project, design, influence, tool, industry
Cluster 5	processes, knowledge, information, system, object, model, support, management, configuration, modelling
Cluster 6	products, companies, customer, functions, development, architecture, collecting, markets, PDP, require

As can be seen from Table 1 the clusters do not represent clear-cut perspectives on a collection and indeed many of the terms that have been used as a basis for clustering are not particularly meaningful within the 'ICED domain'. In this example, in addition to confusing more general and specific themes and subjects within a given cluster, the clustering has also managed to 'confuse' the different perspectives that users might have on a collection of documents. Rather than representing a single coherent viewpoint, a cluster set may instead provide a combination that does not relate to any meaningful topic structure that a human would construct [8]. However, whilst clustering techniques can be relatively poor at discriminating between closely connected sets of documents they are likely to be more effective at distinguishing between clearly delimited collections of documents (e.g. between sales reports and technical engineering documents).

Another important aspect of clustering is that clusters are created on the basis of the available documents at a given instance in time. As the contents of the information collections (on which the clusters are based) changes, so will the clusters. In this respect they are not really suitable for organising incomplete or dynamic collections for browsing – since users will have to learn different organisational structures. However, clustering can be a valuable discovery tool[3] that can be used for the identification of possible classification categories or as a pre-processing step to prepare training collections for automatic classification (see section 5.2).

5. Document classification

Clustering provides an approach for automatically organising large quantities of information where no existing classification schemes exist. However, the resulting grouping of documents will not represent those that a human would choose. In many circumstances where classification schemes already exist, or the investment in developing a scheme is considered worthwhile, automated classification approaches are more appropriate.

[3] As an example within the engineering domain, Matthews *et al* [9] provide an overview and a comparison of the utility of a range of clustering methods for exploring the structure of knowledge in design databases.

5.1 Constraint-based classification

This classification method is conceptually the most straightforward and is reliant on pre-coded sets of *constraints* (i.e. rules or heuristics) that relate the textual content of documents (or more accurately the logical representation of the documents contained in an *index*) to concepts / subject headings in the pre-defined classification schemes. The sets of constraints contain mappings between terms and term phrases, possibly with additional features such as weighting algorithms that allow the combination of multiple sets of constraints [10].

The following provides an example of a simple constraint. Consider a subject category "*Composite Materials*", which might be one of many concept or subject headings contained within a classification scheme. The constraint for this concept could require documents to contain one of the following terms or phrases: *'carbon fibre'*, *'carbon fiber'*, *'glass fibre'*, *'epoxy'*, *'CFC'* or *'GFRP'*. A modified set of constraints could associate constraints with different weightings or require multiple occurrences of single terms or phrases to occur (or even multiple occurrences of multiple terms or phrases) in documents before a document is associated with a concept in the classification scheme.

The advantage of this type of approach is that it is conceptually simple and transparent to system users. The main disadvantage is the manual effort required building and maintaining the constraints that relate words and phrases to the classification categories. Note, however, that to some extent these negative aspects can be addressed by deploying automated means of extracting appropriate keywords and phrases from documents that could be used as constraints [11]. In addition in an engineering context, to some degree sets of constraints are likely to be reusable, since many issues and associated classification schemes (in particular technical and business domains) will be common between companies.

5.2 Machine learning and statistical classification

In machine learning or statistical approaches, users must identify *training sets* of example documents that are representative of each of the pre-defined classification categories. The training process allows the classification algorithm to develop the means to classify documents into particular categories according to the textual contents and characteristics of the training set. Once this training process has been completed the classification algorithm will have been 'programmed' to classify new documents into appropriate categories, according to their textual content. Note that what constitutes a satisfactory training set varies and is highly dependent on the variation of documents for a given application. A wide variety of approaches have been described in the Information Retrieval literature [4][12].

The main advantage of these approaches is that they require little human intervention, once suitable training sets of documents have been identified for the classification categories of interest. The performance of such systems can also be improved over time by incorporating user feedback to continually re-train the network. The main drawback is associated with the identification of the training documents – a far from trivial task in practical situations, especially if the classification scheme is highly sub-divided and structured (although, as previously mentioned, clustering approaches can aid the process of identifying training sets). The 'quality' of the training documents (in terms of how representative the training documents are of those that will be processed by the system) is crucial in determining the classification system performance. As an indication, Letson [5] quotes from a vendor of machine-learning based classification software who notes that, "…the best algorithms available – under optimal conditions, with hundreds of training documents and narrow data sets – can manage 75-80% accuracy in categorisation. On a typical Intranet, with its broad range of content, the best tools, with training sets of 10 or 20 documents, are getting 50-80%

accuracy". In engineering organisations this issues is likely to be compounded in the early stages of a design project by the fact that training documents may not yet exist because they have not been written. Furthermore, in some companies the training sets may be in historic documents that are not available electronically.

Yang and Liu note that "…while the rich literature provides valuable information about individual methods, clear conclusions about cross-method comparisons have been difficult because often the results are not directly comparable" [12]. They go on to present a comparative study of five document classification methods: Support Vector Machines (SVM), a k-Nearest Neighbour (kNN) classifier, an Artificial Neural Network (ANN) approach, the Linear Least Squares Fit (LLSF) mapping and a Naïve Bayes (NB) classifier. They found that the SVM, kNN and LLSF outperformed the ANN and NB approaches for small training sets per category (~10), but all performed similarly when dealing with large training sets. Having noted this, the reader should be aware that these experiments were carried out on a reference collection of newswire reports which are generally short, well-structured documents (i.e. not necessarily characteristic of information that is widely used by engineers).

5.3 Natural language / case-based classification

Systems that use this approach to classify information are reliant on the processing of natural language text to extract document concepts and relate these to nodes in a pre-defined classification scheme. A large amount of research in this area has been carried out [13][14]. Natural Language Processing is typically achieved through a process involving Part-Of-Speech (POS) analysis. This process involves the use of a POS tagging system to analyse the textual content of a document and assign tags to words that reflect their syntactic usage. Note however that due to the complexities of natural language, words can belong to multiple and different syntactic categories in different contexts. For instance, the word *'stress'* can have two quite different meanings. Consider the following two phrases: *"repetitive stress injury (RSI)"* and *"stress concentrations cased by notches"*. In the two examples, the same word *'stress'* is being used in quite distinct technical domains (e.g. ergonomics and stress analysis).

A POS-tagger is therefore required to carry out the following range of processes [15]:

- Identify distinct words (termed *'tokenising'*), a process that is carried out in all full-text indexing tasks.
- Determine the possible meanings in the given context (termed *'morphological classification'*). This is usually implemented in the form or a lexicon lookup table, where words are listed in the lexicon and associated with all of the possible meanings of the word.
- Assign the correct meaning in the given context (termed *'morphological disambiguation'*). There are two main approaches to POS-tag disambiguation, machine learning and rule-based approaches.

Thus, when attempting to classify a newly submitted document, the document is initially linguistically processed into a case representation which highlights the salient concepts and features of a document according to pre-defined mappings between the actual textual content and the case representation. This representation is then processed and similar cases are retrieved from the existing case base. If an identical case match is found then the appropriate classification heading can be extracted. However in the event that an exact match cannot be found then it becomes necessary, using some form of similarity metric, to adapt the new case to those contained within the case base.

A case-based approach has certain advantages particularly that in time, as additional cases are added, then the performance of the classifier will improve. This is particularly likely to apply in an engineering context with a regularly re-used specialist vocabulary. However there are several drawbacks to the approach. It is necessary to spend effort identifying suitable prototypical documents to set up the initial case base. More formidably, the lexical analysis process typically relies on explicitly codified linguistic knowledge to allow the mapping of words into the case representation. This is either a highly manually intensive process or subject to limitations associated with machine learning approaches as previously noted [16].

6. Choosing an information organisation system

The factors that affect the choice of information organisation approach within an engineering organisation are related to wide range of factors. Perhaps most importantly, as indicated in Figure 1, there is a cost / performance trade-off between the different approaches. Clustering provides a fully automatic means for organising large quantities of information, however, in general clusters do not provide as intuitive means of retrieving documents, in contrast to those that are classified into a meaningful classification scheme.

The factors that need to be considered include:

- The existence (or lack) of pre-existing classification schemes and taxonomies – such as organisational breakdowns, Bill of Materials, product breakdown structures, etc.
- The degree of conceptual 'closeness' between documents in a collection (e.g. is it necessary to distinguish between sales and engineering reports (i.e. conceptually not very 'close') or between stress and fatigue analysis documents (conceptually much 'closer')).
- The availability (or lack) of human resources to identify, develop and verify classification schemes.
- The complexity of the overall classification scheme.
- The degree of availability of training sets of documents.
- The minimum acceptable classification accuracy.
- The dynamism of the information collection.
- The number of documents in the information collection.

In the authors' experiences it was found in certain circumstances that automatic classification systems had to be deployed and tested using incomplete and relatively small numbers of documents: for example in the early stages of a project, documents are not available for training purposes, and yet the required classification scheme is known. Furthermore, engineers identified a number of classification categories for any document, including technical issues, product structure issues, commercial and organisational considerations and so on. A typical document could be classified into many categories, and therefore it was difficult to identify training sets with adequate discrimination. In these instances, it was necessary to de-couple the development of classification schemes (and the means of classifying documents into these schemes) from the availability of documents and it was not viable to use machine learning-based automatic classification approaches. These restraints are likely to be common when deploying systems in an industrial setting [17].

Constraint-based approaches have been found to be particularly suitable in applications where documents contain structured fields (e.g. title, keywords, author, etc.) and are written using a relatively restricted vocabulary (e.g. Engineering Change Requests, product catalogues and metadata records). However they are not as suitable as machine learning or statistical classifiers for the general discrimination of longer documents which may relate to multiple subject areas. Clustering requires minimal human input, although the results do not result in intuitive categories, particularly where document collections are conceptually 'close'. However, the results of clustering can provide a useful starting point for the development of classification schemes.

7. Conclusions

Due to the ever-increasing amount of information available within modern engineering organisations, improved approaches need to be developed to allow for the automated organisation of information (since pre-organised information can be more easily retrieved at search time). This paper presents an overview of various *clustering* and *classification* strategies that can be used to automatically organise textual documents.

Document clustering provides a low-cost and fully automatic means of organising information collections, since no pre-defined classification scheme is required. A problem with clustering approaches is that a cluster set may not result in a meaningful topic structure that a human would construct. In contrast, document classification is concerned with the assignment of documents into more intuitive sets of structured categories. However, the construction of such classification schemes is not straightforward and requires extensive expert human input.

Ultimately the factors that affect the choice of information organisation approach depend on the particular application and will be related to a number of factors which include: (i) the existence (or lack) of pre-existing classification schemes and taxonomies – such as organisational breakdowns, Bill of Materials, etc. (ii) the degree of availability of training sets (iii) the availability (or lack) of human resources to identify classification schemes (iv) the minimum acceptable classification accuracy and (v) the dynamic nature of the information collection.

8. Acknowledgements

The funding provided by the UK EPSRC as part of the Innovative Manufacturing Research Centre (grant GR/R67507/01) and the IDEA project (grant GR/L90170) is gratefully acknowledged. The authors would also like to thank the IDEA project's industrial sponsors – Airbus UK, CSC and TRW Aeronautical Systems – for their assistance.

References

[1] Ehrlenspiel K., "Knowledge explosion and its consequences", Proceedings of ICED'97, Tampere, Finland, pp. 477-484.

[2] Lowe A., McMahon C. A., Shah T. and Culley S. J., "A Method for the Study of Information Use Profiles for Design Engineers", Proceedings of ASME DETC 1999, Las Vegas, Nevada, DETC99/DTM-8753.

[3] Foskett A. C., "The Subject Approach to Information", 5th Edition, Library Association Publishing, London, 1996.

[4] Baeza-Yates R. and Ribeiro-Neto B., "Modern Information Retrieval", ACM Press, 1999.

[5] Letson, R. (2001), "Taxonomies put content in context", Transform Magazine, December 2001, http://www.transformmag.com/db_area/archs/2001/12/tfm0112f1.shtml?

[6] van Rijsbergen C. J. "Information Retrieval", Butterworths, 2nd edition, available on-line from: http://www.dcs.gla.ac.uk/Keith/Preface.html, 1979.

[7] Rasmussen E., "Clustering Algorithms", in Frakes W. B. and Baeza-Yates R. (eds.) Information Retrieval Data Structures and Algorithms, Prentice Hall, N. J., 1992.

[8] Hearst M. A., "User Interfaces and Visualisation", in Baeza-Yates R. and Ribeiro-Neto B. (eds.), Modern Information Retrieval, ACM Press, 1999.

[9] Mathews P. C., Langdon, P. M. and Wallace K. M., "New techniques for design knowledge exploration – a comparison of three data grouping approaches", Proceedings of ICED'01, Design Management, Glasgow, UK, 2001, pp. 107-114.

[10] McCune B. P., Tong R. M., Dean J. S., and Shapiro D. G. "RUBRIC: A System for Rule-Based Information Retrieval", IEEE Transaction on Software Engineering, Vol. SE-11(9), 1985.

[11] Gutwin C., Paynter G. W., Witten I. H., Nevill-Manning C. G. and Frank E., "Improving browsing in digital libraries with keyphrase indexes", Decision Support Systems, Vol. 27, 1999, pp. 81-104.

[12] Yang Y. and Liu X, "A re-examination of text categorisation methods", ACM SIGIR Conference on Research and Development in Information Retrieval, University of California, Berkeley August 15-19, 1999, pp. 42-49.

[13] Smeaton A. F., "Progress in the application of natural language processing to information retrieval tasks", The Computer Journal, Vol. 35(3), 1992, pp. 268-278.

[14] Lewis D. and Sparck Jones K., "Natural language processing for information retrieval", Communications of the ACM, Vol. 39(1), 1996, pp. 92-101.

[15] Marcus M., Santorini B. and Marcinkiewicz M. A., "Building a Large Annotated Corpus of English: The Penn Treebank" Computational Linguistics, Vol. 19(2), 1993, pp. 313-330.

[16] Taupin L., "Computer Productivity: Software That Does Your Research For You". Design News, 1999, http://www.manufacturing.net/magazine/dn/archives/1999/dn0920.99/18f1907.htm

[17] Sutton M. J. D., "Document Management for the Enterprise: Principles Techniques and Applications", John Wiley and Sons Inc, New York, 1996.

Chris McMahon*, Alistair Lowe & Steve Culley
Innovative Manufacturing Research Centre
Department of Mechanical Engineering
University of Bath
Bath BA2 7AY
UK
Email: c.a.mcmahon@bath.ac.uk
URL: http://staff.bath.ac.uk/enscam/