Automated Classification of CAAD-related Publications: Conditions for Setting-Up a Keywording System

Bob Martens
Vienna University of Technology
Andre Brown
Universityof Liverpool
Ziga Turk
University of Ljubljana

1 Introduction

In a very direct and astute paper Maver (1995) noted seven principal shortcomings in contemporary CAAD research: what he referred to as “seven deadly sins.” Amongst the major failings that he identified were déjà vu (repeating research already undertaken by others), the failure to validate, and the failure to evaluate. There is such a rich body of research in CAAD that there was, and still is, a clear need to allow researchers ready access to that body of work, so that they can minimize the commitment of those sins. But access alone, though a significant problem, is not enough. What is needed is a way of interrogating the research information in an efficient and effective way. The research entities need to be stored and classified in a way that will allow optimal access, browse and search routines.

One starting point would be to ask, “How does a traditional library handle classification of recorded entities?” Human expertise is used (“manually”), and this is rather cost-intensive as well as time-consuming. As we are dealing with a specific field of knowledge—CAAD—we may assume that those looking for information in a related Digital Library will have some previous knowledge of the domain and thus will achieve a satisfactory search result. Taking into account the fact that thousands of published pages are archived in CUMINCAD, one can hardly expect that this documented scientific knowledge can be handled autonomously. Using advanced search techniques will deliver a generally reliable result, as various fields of the database entries can be investigated simultaneously. Not all records, however, have keywords submitted by the author. The keywords are invented by the authors (rather than being chosen from a list), and their relevance has sometimes to be considered with a little scepticism.

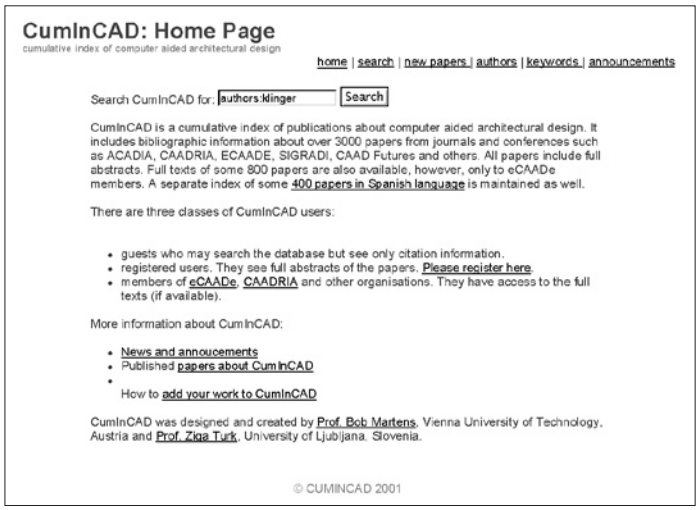

Figure 1. The CUMINCAD interface: entering a search term

The added value of keywording based on a common taxonomy is obvious, even when many other bibliographic fields are searchable. There is a direct relationship between the added value of keywording and the number of searchable documents: the more documents you keep, the more you need keywording. Having each paper tagged with a simple subject allocation cannot be satisfactory in the long term. Subjects need to be refined until they actually reach the precision and consensus of a thesaurus. The continual increase of papers available in CUMINCAD will lead to a poor performance for information retrieval if an effective classification is not undertaken.

Classification by subject specialists is by the far the most effective method, but it is costly in terms of time, and it requires well-qualified people to do it. The question arises as to whether one could achieve a useful result by some automatic procedure based on the text of the title, the summary, or the full text document.

2 Previous work on the indexing and interrogation of CAAD research

The searching of general web-based data and ways of optimizing such search activities has a relatively recent history. Brin and Page (1998) described a very well conceived system for general web searching, called Google, and that system is now in wide general use amongst researchers in all disciplines. Brin and Page noted that at the time they wrote, “despite the importance of search engines on the web, very little academic research has been done on them.”

However, in contrast to general web-based information, the field of research in CAAD represents a relatively well-controlled and homogeneous collection of information. Research on information retrieval systems for more controlled bodies of data goes back to around 1993-4 (Witten et al. 1994), and presents a somewhat different problem in terms of information structuring and searching to general web searching problems. CAAD research has an effective history of less than five decades, and in that time there has been an explosive increase in research and development in the field. However, there are common themes and activities that characterize the research: identifying those themes would enable increased efficiency in interrogation. In order to address this matter, attempts have been made to establish context ontologies, rather than broad global ontologies,

in order to make information retrieval more efficient and effective (Chiang et al. 2001). In the field of CAAD, this kind of approach is likely to prove very useful.

Based on CUMINCAD dataset, Turk et al. (2001) applied text-learning techniques to automatically generate classes of CAAD papers. In spite of the fact that making sense of the categories created is based on intelligent human interpretation, the results were less than satisfactory—they were found very different to human classification of the same works in conference sessions. A similar conclusion has been drawn from a similar analysis of the construction IT papers (Turk 2003).

Chiu et al. (2002) took the research publications of CAADRIA conferences from 1996 to 2001 as the source data with which to develop a data mining system. The system that was developed had the goal of applying "data mining techniques to study the current research paradigms and their relationships based on a research orientated database of collected papers."

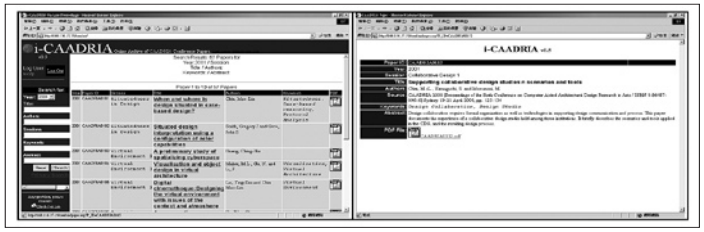

Figure 2. i-CAADRIA

What has come out of this research is a very elegant and enlightening research tool called i-CAADRIA, which has great potential. It is implemented using JAVA, ASP, and HTML, and the data is stored in an ACCESS database. Arising from their analysis, Chiu et al. were able to identify 13 cluster-of-paper domains as follows:

- Design Methods and Models
- Design Cognition
- Collaborative Design
- Information systems
- Design Media and HCI
- Precedents and Prototype
- Shape Studies and knowledge representation
- Generative Systems
- VR and VE
- Simulation
- Prediction and Evaluation
- Regional Information
- Design Education

Obviously these areas are somewhat subjective and are likely to be driven by a particular set of conference themes and culture, given the source of the data. But the areas do provide an interesting foundation for discussion of sub-themes within CAAD research.

In a more detailed study of research themes and sub-domains Martens and Turk (2002) have conducted a detailed analysis of the content and subject areas addressed in CAAD research publications. The work extends and augments the work of Chiu et al. and begins to establish an extensive and informed taxonomy of domain specific terms.

A more focused study by Kosovac et al. (2000) looked at the problem of keywording though the automatic analysis of construction industry documentation. The study focused on a subset of data; that referring to information on roofing of buildings. Concluding from the study, the authors suggested that the "methodology used was found to be highly useful, although it was not sufficient by itself for constructing a thesaurus for the architecture, engineering, construction and facilities management industries [. . .] a considerable human intervention was required." Ciftcioglu and Durmisevic (2001) considered the problem of data mining in the context of construction industry work and have developed a technique that enables data reduction, through clustering, of the large dataset that is produced from an initial analysis.

In the 1980s there was a relatively large body of research undertaken in Artificial Intelligence and Expert Systems. As a necessary part of this research, it was often needed to analyse text based information to provide the database of material with which the systems being developed could interrogate and interact. For instance, Robinson (1981) reports on the use of tree structures to address the problem of retrieving multikey records via range queries form a large, dynamic index in the field of CAD.

It is worth concluding this section by noting the caveats to which some of the researchers in data interrogation techniques have drawn attention. In particular Chiu et al. (2001) note that, with reference to keywords, there are linguistic problems (such as synonymies) as well as the technical problems (devising algorithms). For instance collaboration and cooperation are not exactly the same, but sometimes can be, and are interchanged. Added to that, many authors use key phrases rather than keywords. Are those key phrases used consistently by others? Some are. Ones like *Human Computer Interaction* are established, but how do we identify significant differences, if there are any, between collaborative design, design collaboration, design, and collaboration in keywording? There are also other subtle problems, such as differences in British English and US English. An example would be modeling and modelling, a word that is used frequently in the CAAD domain. These and similar problems require attention when refining the system.

As we mentioned earlier in the paper, the source archive used in this paper is the CUMINCAD system. The creation of CUMINCAD in 1998 was based on the knowledge, that no (web-based) repository existed to support further dissemination of research work in the field of CAAD and Construction IT. Outside Journals, the main channel of exchange has been in the format of annual conference meetings with proceedings as a tangible result. More than 70 CAAD conferences have been organized to date, and many additional conferences have CAAD related papers submitted to them. In total, about 3,200 CAAD conference

(organized by the associations mentioned in this report) papers can be identified in the period 1981-2001. It has to be said that the level of professionalism has increased significantly in the last five years. Also, the process of annual production of papers has been improved substantially, and this has generally happened in tandem with a more rigorous blind reviewing procedure. In 2001, 410 papers were published in conference proceedings. However, this annual production level seems to be a number that will not grow in the near future. The current increase, on a annual basis—without the biannual CAAD futures—is approximately 350 papers. CUMINCAD, which developed on a shoestring budget, supports the important task of information management, as no other similar initiative could be identified so far for this field of research. The associations in charge are established on a non-profit base and have no commercial interest in "making money" from the conference activities.

3 Bibliographic Databases and Aspects of Searching

Bibliographic databases were developed from the traditional library card catalogue in order to enable users to access library documents via various types of bibliographic information, such as title, author, series, or conference date. In addition, these catalogues sometimes contained some form of classification by subject, such as the Universal (or Dewey) Decimal Classification used for books. With the introduction of e-print archives, huge collections of documents in several fields have been made available on the World Wide Web. These developments however have not yet been followed up from a keywording point of view.

3.1 Searching Documents by Subject

Database search engines may offer some features that are designed to improve the precision of the search. Words can be strung together as a phrase, and this phrase searched for. However, searching for a phrase of more than three words is likely to result in a low recall factor because of the flexibility of natural language (particularly English) in representing nuances of meaning by variations in word order. In another approach, limits can be placed on the maximum number of intervening words that are allowed to occur between a pair of chosen words (proximity searches). Of course it is not just the search strategy that counts, but the result of the search strategy when applied to the data. Therefore, exactly which data for a given document are available for searching influences the search result. There are two main uses of a bibliographic database. The first one is to search for a specific item which one already knows about and wants to find out if the digital library has it and, if so, to get access to the document. This is the so-called "referral" approach, a bit like looking up a piece of information in an encyclopaedia. The other main use is when one has a specific problem in mind and wants to find documents that address that problem. It is only with this second type of use that we are concerned in this paper. Basically, this leads to a subject-based approach to the digital library collection.

3.2 Subject connections via references

There is already a system of searching academic literature in a thematic way without any kind of intermediate database. This is via the references to other work or works that have been an accepted and established part of scholarly publications in the subject area. Starting from a core document, one can gradually widen the scope using the references and hopefully arrive at some fairly complete set of relevant documents. The electronic age has again enhanced such an approach without, however, changing it in principle. References in an electronic document can be links to the electronic versions of the documents referred to. The main obvious drawback in this approach is that authors may not have referred to all the relevant material, either due to deliberate omission or just because they do not know about it. Another possible disadvantage is that by definition one can only refer to what already exists at the time of writing a document. From the original document past documents will be reached, but all new documents will be missed. However, this could theoretically be solved through a database of such references, by forward searching from a document to retrieve all other documents that have referred to it later. In practice, the connections via references is not an adequate approach for users needing an exhaustive list of available documents related to a given topic. The process takes too long, and the full coverage is far from guaranteed. The other solution—querying a bibliographic database directly—is potentially much faster, but it may still result in an incomplete result.

At the time of writing, a citation database with nearly 20,000 references (collected from previous CAAD conference papers) has been connected to CUMINCAD.

3.3 Data from the record itself which are available for searching

Here we are only really concerned with data relating to the subject matter of the recorded document, so data like author names do not play a role. Of course, searching for an author can result in retrieving a certain subject, but it is almost never the case that this author is involved with all the documents in that subject area. Traditionally, the title is the item of bibliographic information that expresses the subject content. However, a title is usually far too short to contain a complete description of the subject area in a way that can be used efficiently by a search engine. A specialist reading the title may understand what the document is about, but that specialist is using all sorts of prior knowledge into which context the new title can be. Therefore the recall factor of a title-based search is likely to be low.

Sometimes titles are misleading. They can be guilty of overestimation concerning the real content, or the title simply promises too much. Furthermore, as the number of documents

in the database steadily increases with time, the precision of title-based searches is likely to decrease as well. An extension of this, which has become much easier to realise for electronic documents, is that more of the text than just the title can be used for searching. Ideally this would relate to an extended abstract that summarises all aspects of the work covered in the paper, but it still remains that it is very time consuming to read all this information.

3.4 Subject Indexing

The grouping of source material by subject is called subject indexing or keyword enhancement. When we say "keyword," this could of course mean a phrase of two or more words. There are two very different ways of doing this: to choose terms from a fixed thesaurus or to use free keywords that can be chosen by the indexer at will. The strategy of assigning keywords will obviously depend on which parts of the document itself (title, abstract, full text) are also available for searching. It has to be stated that CUMINCAD has been focusing on conference papers; in many cases the supplementary addition of keywords simply did not take place.

Allocating keywords on a free basis could also use terms that are not present in the document, but in practice this technique is mainly used for adding useful words or phrases taken from the text, such as section headings and other specific words that could help in improving the recall factor of the search. Free keywords can also be useful for indexing terms containing special characters that would not be completely recognised if they appeared in the title or abstract. Free keywords can also be a useful way of adding synonyms of terms that appear in the text. But it would be better in general to handle synonyms at the search input end rather than adding them to each record when they occur.

The efficient allocation of keywords from a fixed thesaurus makes the most demands on the indexer, as the documents have to be well understood. The indexed terms may not appear in the same way in the text at all, which can give this method a big advantage over any strategy that just uses the text of the document. Of course, such a method requires the existence of a complete, precise, and up-to-date thesaurus, which is quite difficult to achieve in a rapidly-changing specialized research area like CAAD.

In practice, these two forms of indexing are extreme cases. Real approaches have aspects of both, even though they may be closer to one than the other. Thus, the drawback of having a fixed thesaurus is that the thesaurus itself has to be modified to keep up with developments in the field. On the other hand, free keywording can be chosen to conform to a minimum set of rules, instead of being completely free and just taking the words as they appear. For example, it could be decided to choose singular forms instead of plurals. In fact, after a period of use, listing the terms which have been given as free keywords does give an empirical thesaurus, that can then be used to standardize the keywords which are subsequently assigned, in order to improve consistency.

4 Beta version of a System for establishing Topics and Associated Keywords

The reasons specified above encouraged us to take a closer look at the possibilities of setting up a system of keyword categories for the area of Computer Aided Architectural Design. The range of tasks was limited to three key measures:

(i) compilation of a list of no more than 30 primary topics (key categories) within the domain
(ii) definition of 5-10 subtopics each per keyword category, and
(iii) selection of up to 10 relevant typical papers from the database associated with these topics.

Given the existing CUMNICAD repository as the source, first of all a provisional list was established based on the frequency of entries in the CUMINCAD-database field "keywords":

- 2D Representation
- 3D City Modeling
- 3D Modeling
- Animation
- Artificial Intelligence
- Case Based Reasoning
- Collaborative Design
- Communication
- Computer Integrated Construction
- Constraint Based Design
- Database Systems
- Design Methodology
- Design Process
- Digital Design Education
- Digital Media
- Environmental Simulation
- Generative Design
- Human-Computer Interaction
- Image Processing
- Interactive Design
- Knowledge Modeling (KM)
- Learning Environment
- Object Oriented Modeling
- Performance Simulation
- Shape Grammars
- Virtual Design Studio
- Virtual Environments
- Virtual Reality
- Visualization
- Web Design

For the definition of related subtopics and typical publications expert-help within the CAAD-community was requested. In order to avoid additional workload, the experts did not see the whole set of keywords, but only one category. As this is work in progress, the inquiry is not yet completely finished. To give an impression of the result one example is displayed here:

- Collaborative Design
  - Collaborative Teamwork
  - Collective Authorship
  - Computer Supported Collaborative Design
  - Computer Supported Collaborative Work
  - Distributed Modeling
  - Distributed Workgroups
  - Groupware
  - Groupwork
  - Information Sharing
  - Multi-User Workspace

The provisional list is covering a vast amount of information in order to feed the bibliographic database. Naturally a specialized field like CAAD does not consist of fixed topics, and thus the systematics introduced above are to be adjusted in line with current developments. On completion of a “Betaversion 1.0” tests with clustering processes can be initiated. The user community might be particularly interested in hints as to comparable publications.

Even in a list like the one above we see potential problems that are outlined in Section 2. There are terms such as “distributed modeling,” where there is potential ambiguity and confusion on at least two counts: firstly the British and US spellings of modeling/ modelling, and secondly the use in a phrase and singly as just “modelling.”

Also there may be overlap across subtopic groups. For instance, distributed modelling can be a subtopic of Collaborative design. It could easily be a subtopic of a different primary keyword: modelling. Hence the generic structure that is identified for keywords and subtopics is not a pure tree structure; the branches of the tree intermesh. Hence the idea of a rationalised and strict hierarchy of terms is neither realistic nor possible.

At least at present, it is not reasonable to consider that we could impose a lexicon of standard terms on the CAAD research community. New terms will be added by researchers, and other terms will become unused or redundant as the research field develops. One of the interesting aspects of the research reported here is to track and reflect that evolution; so imposing a lexicon would be unduly restrictive and artificial. The problem, then, is one of dynamically tracking the evolution, but also we hope that there will be scope for introducing the idea of commonly used core terms in a standard form so that we can all search for research in a particular sub-topic area.

5 Text Learning and Clustering

The result of the effort of human experts is a list of 30 or so topics, each with about ten subtopics and about 10 related papers. It can be seen that for the sample groups of around 300 papers (out of over 5000) we have a manually defined the topic to which a paper belongs. We also have lists of 300 keywords, each associated with one topic. This data will be used to classify the 5000 papers in two ways:

1. keywords related to the topics will be used as learning set to train a classification tool. Based on this training, the 5000 publications will be classified.
2. papers, identified as typical for a topic, will be used as a training set to train a classification tool.

The results of the classification will then be combined and compared. Human experts will be asked to review the resulting classes and:

1. Identify papers also typical for a topic.
2. Identify papers clearly classified into a wrong topic.

Based on these comments, and a review of the effectiveness of performance of the system, the classification engine will be trained again. After a couple of iterative cycles the categories should be quite well established. The classification engine and the system to suggest a different classification will be part of an on-line knowledge management service which is part of the SciX system and which has the goal of providing technologies for low maintenance efforts and self-organisations of digital libraries such as CUMINCAD. After the bulk of the papers have been categorised it is envisaged that each new paper will be classified by the system automatically.

6 Conclusions

The number of recorded publications in CUMINCAD has grown year by year and growth has reached a steady state. A need for setting up a framework concerning the classification of stored knowledge in this extensive Digital Library has now become clear, and this fact defined the starting point of the work reported here. The procedure for gathering keywords and related sets of subtopics has been described in this paper, and the clear distinction is made between this activity and the implementation of an automated classification system, which is planned to follow at a later stage within the CUMINCAD development process. It will be interesting to see how efficiently the automated process that is planned for the next stage of development, will prove to be. An irony occurs to us. That is, that in this paper in indexing and searching data on CAAD research we may well have missed referencing, and drawing on, work by others on the subject of indexing and searching. Apologies are due if this is the case, but also it is reassuring that the CUMINCAD system was very useful in tracking down certain relevant work in this area.

Acknowledgements

The topics, associated keywords, and related typical papers were compiled in the framework of the Scix-Project. Invaluable input in the course of development was made by the eCAADe-council and individual experts, to whom we extend our sincere thanks.

References

Brin, S., and L. Page. (1998). *The anatomy of a large-scale hypertextual web search engine*. Stanford, CA: Computer Science Department, Stanford University.

Chiu, M. L., C. J. Lin, T. S. Jeng, and C. H. Lee. (2002). Re-searching the research problems with CAAD: Datamining in i-CAADRIA. *CAADRIA 2002: Proceedings of the 7th International conference on computer aided architectural design research in asia cyberjaya (Malaysia).* 18–20 April 2002, pp. 031-38.

Chiang, C., and Story. (2001). A smart web query method for semantic retrieval of web data. *Data & knowledge engineering*, Elsevier, 28:63-84.

Ciftcioglu, Ö., and S. Durmisevic. (2001). Knowledge management by information mining, *Proceedings of the ninth international conference on computer aided architectural design futures*,533-545,Eindhoven 8-11 July 2001.

Kosovac, B., D. J. Vanier, and T. M. Froese. (2000). Use of keyphrase extraction software for creation of an AEC/FM thesaurus. *ITcon* 5.

Maver, T. W. (1995). CAAD's seven deadly sins. In *Sixth International Conference on Computer-Aided Architectural Design Futures* Singapore,21-22, 24-26 September 1995.

Robinson, J. T. (1981). The K-D-B-tree: A search structure of large multidimensional dynamic indexes. *CADline* 22.

Turk, Z., T. Cerovsek, and B. Martens. (2001). The topics of CAAD: A machine's perspective. In *Proceedings of the Ninth international conference on computer aided architectural design futures,*547-560,Eindhoven. 8-11 July 2001.

Turk, Z., and T. Cerovsek. (2003). Mapping the W78 papers onto the construction informatics topic map. *Proceedings of the 20th CIB W78 conference on information technology in construction.* Waiheke Island, Auckland, New Zealand. 23-25 April 2003 (forthcoming).

Witten, I., A. Moffat, and T. Bell. (1994). *Managing gigabytes: Compressing and indexing documents and images*. New York: Van Nostrand Reinhold.