# RE-SEARCHING THE RESEARCH PROBLEMS IN CAAD

*Data Mining in i-CAADRIA*

MAO-LIN CHIU*, CHIEH-JEN LIN**
**Department of Architecture, National Cheng Kung University*
*No. 1, University Road, Tainan 701, Taiwan*
***Department of Interior Design, Tainan Woman's College of Arts & Technology, Yung-Kang City, Tainan, Taiwan*
*mc2p@mail.ncku.edu.tw, jackyljz@seed.net.tw*

AND

TAY-SHENG JENG*, CHIA-HSUN LEE*
*tsjeng@mail.ncku.edu.tw, n7689103@dec4000.cc.ncku.edu.tw*

**Abstract.** This study attempts to develop an online CAAD research archive of conference papers, i-CAADRIA, and apply data mining techniques to find research patterns. Research papers are clustered for building semantic relationship. The system and early feedbacks are presented. This study suggests that smart web query and user interface can enhance our understanding of the research patterns.

## 1. CAAD Research

The research problems faced in the academic society are continuously expanding and varied. Understanding the previous research efforts is an important task in the development of computer-aided architectural design (CAAD). Our preliminary study found that researchers spend more time on formulating their research problems, and undertake the steps of reviewing literatures, constructing research methodologies, applying analytic or computational techniques, implementing prototype systems or conducting experiments, and finally concluding the research result based upon the previous findings. Research papers such as conference or journal papers serve as the foundation of research and often provide the impetus for further research. Consequently, researchers continuously re-search the same problems with various approaches or techniques.

The objective of this paper is to develop an online CAAD research archive of conference papers and examine the research patterns by novel information mining techniques, i.e., apply data mining techniques to study the current research paradigms and their relationships based on a research-oriented database of collected papers. This study serves both the methodological and technical purposes, i.e. to search the linkages within CAAD research paradigm and study the mining techniques.

## 2. i-CAADRIA: An Online Research Archive

### 2.1. DATASET AND KEY QUESTIONS

Currently, we have a collection of 283 research papers accumulated in the last six CAADRIA conferences from 1996 to 2001. These contain the efforts from various disciplines. The original dataset has only paper titles, authors, abstract, full text, and without keywords. A concise index file is maintained for organizing these files. The papers appeared in the proceedings are digitized and transferred into pdf format, and each paper is assigned with a unique id. Figure 1 demonstrates the paper contents on the right side and the research process on the left side, while both are related. From the research point of view, re-searching the research problems in CAAD help us understanding the trend of research topics, the problems and the approaches, the contributions, and the distribution of research units.

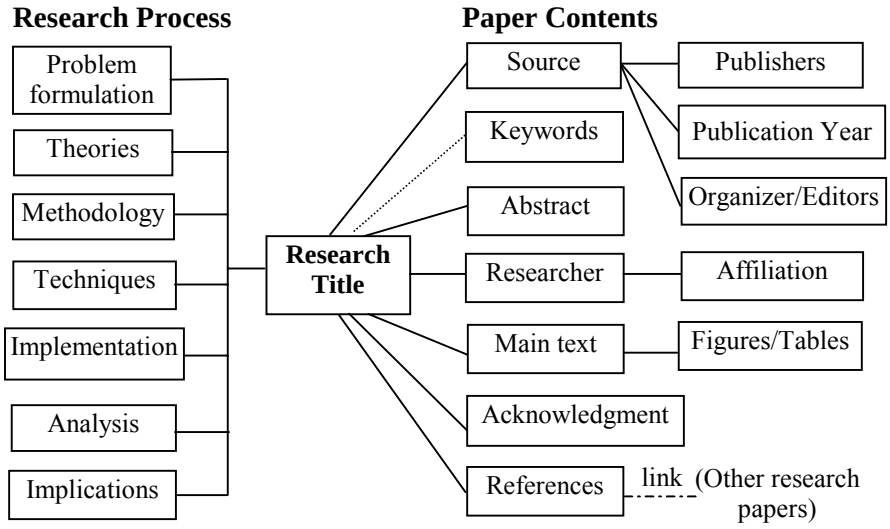


*Figure 1.* Research Process and Paper Contents

Furthermore, knowledge management is important in the research field for continuous development. A 6W (what, who, when, where, why, and how) approach is applied to analyze the research contents, including questions such as: What are the research problems in CAAD and who are involved in these researches? How are these problem solved? What kinds of methodologies or research paradigm are involved? What are the applications? Why are the applications important or fail? Therefore, we need an information mining mechanism to answer the above questions.

### 2.2. ONLINE ARCHIVE

To fulfill the research requirement, we first implement an online system, i-CAADRIA, by JAVA, ASP, and HTML to integrate with the ACCESS database. The primary function of online systems is the repository of research papers. Figure 2 demonstrates the snapshots of i-CAADRIA. CUMINCAD (CUMulative INdex on CAD) provides an example that an online database can be used as the data mining source for research [Martens and Turk, 1999]. Similarly, i-CAADRIA is an online archive on the web for sharing the resource by an accumulative process of collecting new inputs. In addition, the system is enhanced for mapping the research issues by the 6W scheme with analytic interfaces. Whether the research patterns that can be mined are represented in terms of a structure that can be examined, reasoned about, and used to inform future decisions.

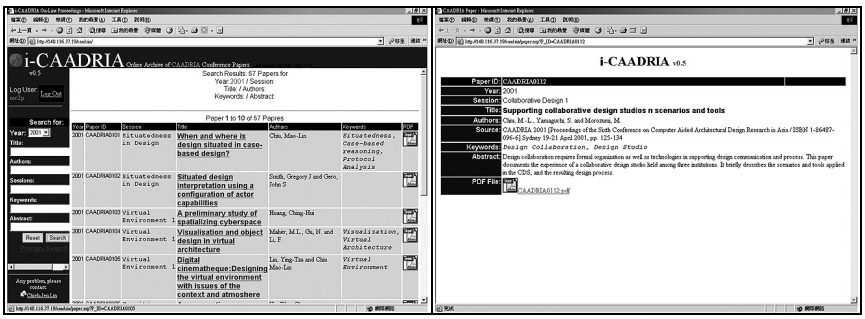


*Figure 2.* Snapshots of i-CAADRIA

## 3. Data Mining in i-CAADRIA

Data mining is defined as the process of discovering patterns in data [Witten and Frank, 2000]. In this case, the patterns are research patterns. Clustering is one of the essential methodologies in data mining to deal with a large dataset in the process of data reduction [Ciftcioglu and Durmisevic, 2001].

### 3.1. DATASETS

In general, related papers are classified into research domains in accordance with their relevance, Table 1. The possible correlations are among papers in terms of the research domains, techniques, or researchers. In terms of research paradigm or domains, we found that most researchers continuously contribute to the fields of collaborative design, shape semantics and knowledge representation, virtual reality and virtual environment (VE), and design education. The conference sessions are in fact the basic clusters of paper domains by organizers, while the classification is often subjective for the organizational convenience.

TABLE 1. Summary of CAADRIA Paper Research Domains from 1996 to 2001

| Areas* / Year | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 |
|---|---|---|---|---|---|---|
| Design methods and models | 6 | 2 | | | | 6 |
| Design cognition | 2 | 1 | | | | 6 |
| **Collaborative design** | 3 | 8 | 5 | 6 | 4 | 8 |
| Information system | 1 | 2 | 7 | 9 | 8 | |
| Digital media and human computer interaction | 1 | 3 | 3 | 8 | 12 | |
| Precedents and prototype | 1 | 4 | 4 | 3 | | |
| **Shape studies and knowledge representation** | 2 | 7 | 10 | 3 | 4 | 8 |
| Generative systems | 4 | 4 | 1 | 2 | | 3 |
| **Virtual reality and VE** | 1 | 4 | 7 | 5 | 6 | 12 |
| Simulation | 1 | 1 | | | 6 | 3 |
| Prediction and evaluation | 2 | 7 | 5 | 4 | | |
| Regional information | | | 3 | | | 2 |
| **Design education** | 8 | 2 | 3 | 3 | 8 | 8 |
| Total | 31 | 45 | 48 | 43 | 48 | 56 |

* Some areas are renamed and reclassified from the original sessions for comparison purposes.

### 3.2. CLUSTER ANALYSIS

The CAADRIA dataset is broadly divided into two categories as (1) clustering with supervised and (2) clustering with unsupervised learning. In the supervised sub-dataset, we are concerned with a collection of labeled data that come in the form of ordered pairs. This can be seen as a feature value describing the data and its class assignment. In the unsupervised sub-dataset, we start with the number of clusters assigning each pattern to a separate cluster and proceed to merge the clusters that are the closest with the distance function. Eventually, the association between the feature values and the class assignment are endeavored for determinations.

Because the research domains, analytic methods, and technologies applied are correlated with timing, the steps in data mining are undertaken as follows: (1) dividing all papers into sub-datasets in according with years, Figure 3, (2) choosing typical sample files in association with particular domains or subjects by experts, (3) scanning paper abstracts to collect keywords for building semantic relationship, (4) using the semantic relationship to analyze the entire dataset for classifying or associating relationships, and (5) evaluating and redefining the semantic relationship. For example, we start with the sample files in the field of collaborative design to study their relations in the data of year 2001, and examine the domain backward. The keywords are selected based on the frequency of occurrence in their abstracts.

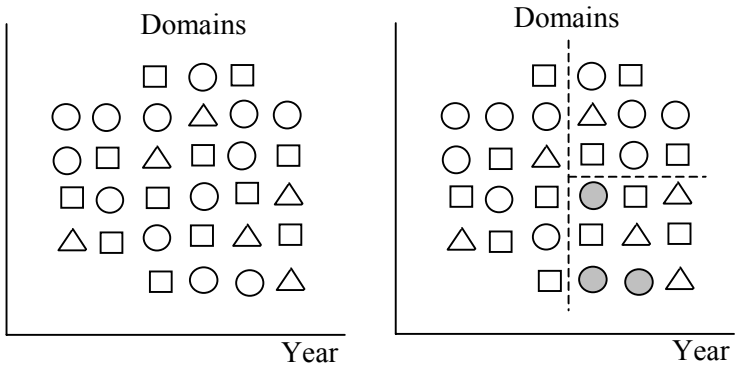


*Figure 3.* Subdivision of Dataset by years and domains

### 3.3. THE SEARCH MECHANISM

Based on literature surveys, approaches to web search include (1) syntactic search, (2) metadata search, (3) query by examples, and (4) navigational search. Currently, i-CAADRIA is implemented with three levels of search functions such as (1) general search (year, keyword), (2) detailed search (year, title, researcher, affiliation, and references), and (3) advanced search (association of features). While key information such as keywords derived from abstracts is accumulated, there are linguistics problems (such as synonymies) as well as technical problems (algorithm) in classification.

Although some attempts have been made to develop universal ontologies, context ontologies are more effective for the semantic retrieval of web data [Chiang, et al., 2001]. Therefore, we adopted the smart web query (SWQ) method proposed by Chiang, Chua and Story (2001) for exploring domain semantics (i.e., knowledge and information), and context ontologies in

selected domains are further analyzed. Figure 4 demonstrates the search engine and its components.

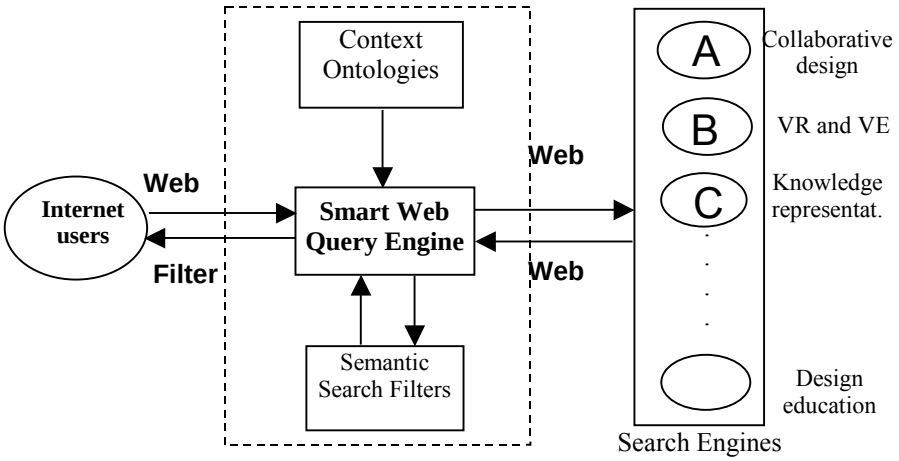


*Figure 4.* Search engine and its components

## 3.4. USER INTERFACE

Traditional information management systems use folders and focus on separating information that creates barriers between information. The result of search in dataset is often listed with detailed and separated information. On the contrary, links enable users to connect information into a network of logical association. Furthermore, visualization of association becomes more feasible and important on the web searches now [Inxight]. i-CAADRIA is implemented by the TheBrain program to visualize the associations of certain features. Figure 5 demonstrates the associations in design education (left) and collaborative design (right). Therefore, users can search or trace desired research contents through iterations of searches by selected inputs.

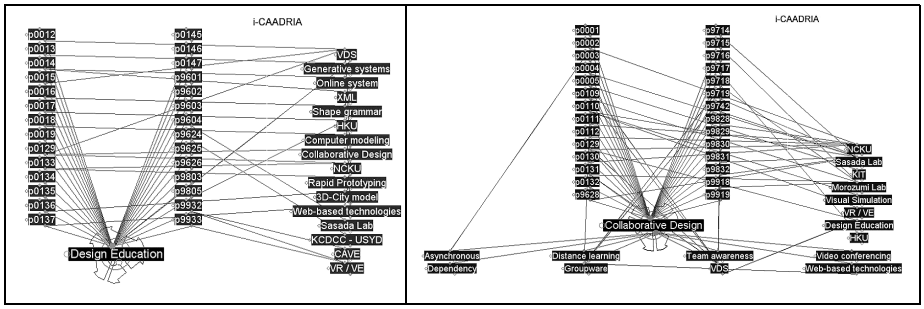


*Figure 5.* Association of Papers in Collaborative Design (left) and Design Education (right)

## 4. Discussion

### 4.1. MACHINE VS. HUMAN VIEWPOINT

We have compared the results of different searches. The size of information is further tested starting from 50 up to 300 papers. We found same terminologies or synonymies in the same domain. While a list of keywords is generated from the frequency of occurrences in paper abstracts, we found that there are strong and weak associations among these keywords in the dataset. Only the strong associations are useful for cluster analysis. When the number and scope of search is gradually increased, the effectiveness of association or search is improved. However, there are also limitations in data mining from the machine viewpoints [Turk and Martens, 2001]. The limitations are evident as the intelligent and learning ability. The study then attempts to enhance the search function of i-CAADRIA by a machine learning approach from both the user behaviors and research needs. On one hand, we need experts in each domain to identify the key papers for information mining. On the other hand, we need to record the user behaviors on the log files to understand their interests. Further analysis is undertaken to classify the search patterns by frequencies and association.

### 4.2. KNOWLEDGE DISCOVERY

We also discovered certain knowledge from the research domain and researchers about the CAAD applications. For instances, many researches in design education are applying various techniques such as virtual reality, collaborative design, web-based technologies, shape semantics, computer modeling, or visualization to CAAD. The preliminary findings demonstrate that the continuity of same research subjects or groups presents the research interests during the time span. Therefore, the relationships between time, subjects, and researchers are considered key attributes for association in data mining. Meanwhile, references are also important sources for research in addition to the main text. From the searches, we can detect the important references that continuously appear in certain research domains. For instance, there are 569 references abstracted from 56 papers in 2001. However, 19 references are overlapped and quoted by different researchers.

### 4.3. USER RESPONSE

We have found that different users are interested in different features in the uses of i-CAADRIA. From the user points of view, their searches are driven for different purposes such as: (1) search for definition and applications, (2) search for the researchers and their affiliation, and (3) search for references.

Meanwhile, the search behaviors can be distinguished by their research experience. Young researchers may search to gain further understanding of the fields, while experienced researchers tend to search for specific interests by ad hoc manners.

## 5. Conclusion

We have constructed the online research archive, i-CAADRIA, and collect 6-year conference papers. The findings in this study include the research patterns and data-mining techniques. While the dataset is small, some research patterns are revealed. Associative reasoning is useful for searching key information from the user point of views, while different user behaviors exist. This study suggests that smart web query and user interface can enhance our understanding of the research patterns and approaches. These findings provide the foundation for future development.

Furthermore, architecture research as well as design involves a large-volume data management activity. Therefore, an association-based information management mechanism is needed for research. Our future works is implementing the data mining techniques applied in i-CAADRIA to support distributed research or design collaboration.

## Acknowledgements

The authors are grateful for the provision of proceedings from the previous conference organizers in the last six CAADRIA host institutions (1996 HKU, 1997 NCTU, 1998 Osaka U., 1999 TJU, 2000 NUS, 2001 USYD).

## References

Chiang, Chua and Story: 2001, A smart web query method for semantic retrieval of web data, *Data & Knowledge Engingeering* 28, 63-84, Elsevier

Ciftcioglu, O. and Durmisevic, S.: 2001, Knowledge management by information mining, P*roceedings of CAAD Futures 2001*, Edinhoven, Netherlands, 533-545

CUMINCAD, url at http://itc.fgg.uni-lj.si/cumincad/

Inxight, url at http://www.inxight.com/

Martens, B. and Z. Turk: 1999, Working Experiences with a Cumulative Index on CAD: "CUMINCAD", *Proceedings of the eCAADe Conference*, Liverpool, UK, 327-333

THEBRAIN, url at http://www.thebrain.com/

Turk, Z., and B. Martens: 2001, The Topics of CAAD - A Machine's Perspective, *Proc. Of CAAD Futures 2001 Conference*, Edinhoven, Netherlands

Witten, I. H. and E. Frank: 2000, *Data Mining – Practical Machine Learning Tools and Techniques with JAVA Implementations*, Morgan Kaufmann