Smart Semantic Query of Design Information in a Case Library

Chieh-Jen LIN and Mao-Lin CHIU

1 Department of Interior Design, Tainan Woman’s College of Art & Technology, Taiwan
2 Department of Architecture, National Cheng Kung University, Taiwan

Keywords: associative reasoning, case-based reasoning, data mining, design case, semantic

Abstract: This paper is aimed to establish a smart query mechanism based on the semantic relationship among design cases in a case library, Case Base for Architecture (CBA). A case is composed with three levels of design information: “general”, “analytical” and “recommendation”. The data mining technique is applied to analyse the case information of CBA and extract a list of keywords of design concepts and knowledge from the analytic information and recommendation. By clustering and ranking the semantic relations of those keywords on the strength of machine and addressing them to selected design cases, we can organise the semantic relations among cases. Based on the results, we build a primary query interface to help users to retrieve and inspire their associative reasoning for design thinking. The semantic query mechanism is presented and discussed.

1 INTRODUCTION

1.1 The Case-Base for Architecture

Case-based design and reasoning is an important research paradigm in design studies. Design cases are often becoming the impetus for design thinking and problem formulation based on analogy. The critical function of a case-based system is the searching and retrieval engine of cases. The associative reasoning based on semantic relations among cases is the primary approach for design case searching and learning. Therefore, semantic relations are critical for users to retrieve cases and perform design reasoning.

The objective of this paper is therefore to establish a smart query mechanism based on the semantic relationship among cases. This research is on the foundation of a case library, “Case Base for Architecture” (CBA) of office building cases implemented on the web, as a case repository and a learning environment for case-
Digital Design

Based design. For the further research of case-based design and reasoning, we have extended the cases of CBA from office buildings to house cases.

1.2 Three Levels of Design Information

Design is often goal-oriented, particularly for problem solving. A case-based system should store the design context of cases in terms of features, graphics, and texts that enable users to associate with or reason about their design contexts or situations, and then help users to retrieve the most appropriate cases in the system as the full or partial solutions. Meanwhile, design information needed to answer the problems has different aspects of design knowledge, and can also be classified into three levels of information, namely “general”, “analytic” and “recommendation”, Figure 1.

![Figure 1 Abstraction of Case Contents in CBA](image)

1) The first level is the general information of cases that includes the most explicit attributes of case features such as participants, events and time factors, site factors, building factors (usage, building type, style, area, stories, height, structural system, etc.) and so on. Essentially, the information of this level is represented in the “attribute-value pairs” format in the CBA system similar as other case libraries.

2) The second level is the analytic information of cases. The ideal analytic information should be the essence of concise design knowledge extracted from cases that can inspire further design thinking. However, the information of this level usually involves many non-text media materials such as images, drawings and diagrams. Since non-text materials are difficult to directly convert into an indexical and searchable format, we will alternatively encode the information by attaching annotations, such as the structural prototype or other graphic diagrams, to represent the feature of those non-text materials.

3) The third level is the “recommendation” of cases that is the user feedbacks or commentaries of cases from case creators, the professional critics or the users of the system. The contents of the recommendation information
usually are texts and contain more conceptual and deeper degrees of design knowledge. The recommendation information involves useful design experience and knowledge for users, but need more assistance to help users to index and retrieve the suitable information.

1.3 Keyword Searches of Design Information

At different design stages, users request different information for various needs (Chiu et al. 2002). Different searching methods and functions are required in response to these intentions. Although the graphic-based search modality, such as retrieving design cases by diagram sketches, may have more convenient and efficiency to retrieve graphic information (Gross 1995). However, this paper focuses on the text-based search methods, and we will examine feature-based and keyword-based search methods.

The ordinary searching method in database usually relies on “keywords”. But the keyword-based search has two major problems that lead to the inefficiency in the retrieval of useful cases. First, the inaccuracy between the query input by users and the data stored in the system. Secondly, the ill-structured information in system may cause the useless results of searching. Despite the exist problems, keyword search is still the easiest way to identify and represent a abstract concept, and a “keyword” actually can be recognized as a container and an index of abstract concepts and design knowledge. Therefore, the keyword-based search modality still has the significance in general search and the edification of associative reasoning, but it needs the improvement of the effectiveness in search. A promising way to overcome the difficulties is to employ the data mining techniques to the case library in retrieving appropriate cases to satisfy design intentions by establishing the semantic relations among cases.

2 DATA MINING IN CASE LIBRARY

Data mining (DM) is a technique and processes for discovering and describing structural patterns in data (Witten et al. 2000). DM defines the overall process of extracting the implicit information of a data source and converts it into explicit form. In order to extract the implicit correlations among cases, we apply the data mining techniques to analyse the information in CBA. Firstly, we extract a lexicon of terms and apply the domain knowledge of design to filter out the useful keywords. Secondly, we establish the semantic relations of those keywords by means of the statistical co-occurrence of the context by machine automatically. Thirdly, we build the inverted-index of cases by the terms of those keywords, and draw down the relation and similarity among cases based on the semantic relations of those keywords. Finally, we build a smart query interface for CBA to provide assistance for searching, retrieval and presentation of semantic relationship among cases.
2.1 Extraction and Ranking of Keywords

Two primeval lists of terms are extracted from the analytic and recommendation information of CBA based on the Zipf’s distributions (Baeza-Yates et al. 1999). Then we rank those terms by means of the tf-idf scheme of term-weighting strategy. (Baeza-Yates et al. 1999) Basically, the number of raw terms extracted from case by machine automatically is about one third of words of case information. For example, the house case “Villa Savoye” designed by Le Corbusier contains about 900 words, composing of three kinds of analytic information and one recommendation, can be extracted about 300 distinct useful terms.

However, the raw terms extracted automatically are not all the useful keywords for retrieving case or presenting the information of case. Therefore, we raise the statistic threshold and apply the domain knowledge with expertise to filter out the most important keywords to build our keyword lexicon. Consequently, we obtain about distinct 40 keywords for the case “Villa Savoye” and average 30 keywords for a single-family house case in CBA.

2.2 The Semantic Relations of Keywords

Since the similarly semantic words should have the similar contexts, then we can gather the statistics of contextual of keywords to and convert the raw frequencies of contextual words into the tf-idf weight, then arrange those weights into a context vector $\vec{v}_i$. The similarity $\text{sim}(k_i,k_j)$ of two keywords can be quantified by the cosine of the angle between two context vectors. (Equation 1) Therefore, we can rank the semantic similarity of keywords and connect the semantic relations among keyword in terms of contextual words of keywords. (Equation 2)

$$\text{sim}(k_i,k_j) = \frac{\vec{v}_i \cdot \vec{v}_j}{|\vec{v}_i| \times |\vec{v}_j|} \quad (1)$$

$$x_{ij} = \frac{\sum f_{ij} \times f_{ij}}{\sum (f_{ij} + f_{ij} - (f_{ij} \times f_{ij})) \times x_{ij}} \quad (2)$$

In order to establish the semantic relations among cases, we first apply two methods to cluster keywords: (1) the association clustering and (2) the dependence clustering. Since the co-occurred frequencies $(f_{ij}, f_{ij})$ of keywords $(k_i, k_j)$ in a same case $(e_i)$ usually have a semantic association; the association clustering is based on the statistical co-occurrence of keywords inside a same case. The dependence clustering of cases is based on the distance function $r_{ij}(k_i,k_j)$ of the keywords co-occurring in a same case. If two keywords occur in the difference sources of information, then the clustering factor of them should be zero. (Equation 2) Based on the values of the clustering factors, we can rank the correlations among keywords and expand the queries from users and retrieve more relative cases based on those relative keywords.
2.3 Clustering of Design Cases

The idea of “keyword vectors” provides a probable clustering method to avoid that contradictory between the explication and the elasticity by the strength of machine. Based on the previous process, the keywords and its frequency vary from case to case, but should have some patterns for the similar type of cases. Therefore every case $c_j$ can be expressed the $tf-idf$ weight of its keywords into a vector $\vec{v}_j$. Therefore all case has a $t$-dimensions of keyword vector to represent its features and $t$ is the total number of keywords of all cases. For example, three house case: “Villa Savoy”, “Gehry House” and “Koshino House” have about 75 distinct keywords, then the keyword space is 75-dimensions. Then we can locate every case in this $t$-dimensions space and further cluster them by calculating the distance and the cosine similarity among cases in this space. Figure 2 demonstrates the probable distribution of cases before and after clustering by two selected keywords such as “function” and “form”.

![Figure 2](image)

**Figure 2** Probable distribution of cases before (left) and after (right) clustering by two selected keywords such as “function” and “form”.

2.4 Discovery of the Semantic Relations among Cases

After the clustering and ranking the semantic relations among cases and their keywords, we can build the inverted-index of cases based on the ranking and improve the efficiency of searching. Once mapping the “keywords” inverted to cases, the semantic relations of “keywords” actually unfold the semantic relations among cases. Furthermore, we can discover the patterns of cases by selecting a collection of most relational “keywords” based on the ranking, and then apply the “keyword vectors” algorithm to cluster and classify the cases automatically.

Because of potential usability, we currently only collect two building types of cases: office buildings and single-family houses and test single-family houses first. We found that the number of keywords is not linearly increased with the number of cases. For example, there are average about 30 keywords for a low-rise house case in CBA, but only about 75 distinct keywords for three cases. Similarly the number of cases clustering will not linearly increase with the number of cases. Therefore, we can represent the patterns and the semantic relations of cases with those keywords.
3 SYSTEM IMPLEMENTATION

3.1 Framework and Interface

Figure 3 demonstrates the framework of CBA system. After the pre-processing of database, the system has established a lexicon of keywords, and then built up the ontology of semantic relationships among cases by means of ranking and clustering the semantic relations of keywords. Once some queries are inputted into the system, the smart query mechanism of CBA will retrieve the relevant keywords from the lexicon of keywords. Then, the CBA will filter out the most relevant cases and the ranking based on the semantic relations of relevant keywords by the ranking factors stored in the ontology of semantic relations.

By applying the techniques of JAVA, ASP, HTML and Flash-based graphic interface to integrate with the Microsoft Access database, we implemented the CBA system interface on the web. For the effective representation of relational ranking, we simplified those complex network graphics interfaces to a multi-columns table to reveal the relevant keywords and the ranking of relational intensity. Figure 4 demonstrates the snapshots of CBA.

Figure 3 Framework of CBA

Figure 4 CBA Interface - Queries (left) and Retrieved Cases Information (right)
3.2 Smart Semantic Queries and Search Mechanism

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., design knowledge), and context ontologies in selected domains are further analysed. Because the keywords are generally context-sensitive, the “function, behaviour, structure” (FBS) model is applied to classify keywords and assist the matches among inputs and outputs, or queries and results (Maher et al. 1995).

According to the prior discovery of patterns, including “keyword vectors”, association and dependence clustering factors of keywords and cases, the CBA system currently implements three levels of search functions such as (1) the general search by feature matching and full-text keywords matching, (2) the smart keyword search by comparing the query form users with the keywords list stored in system, and (3) the keyword browser that allows users freely surf the entire keyword list and the relational network of keywords to retrieve cases that users are interested. The interfaces as demonstrated in Figure 4 are used for queries and retrieved case in CBA. The patterns of clustering are shown as lists of keywords with the ranking of correlations. Once a query input by users or picked up from keyword lists, the relevant keywords or cases and the ranking of their relations emerge on the interface.

3.3 Preliminary Evaluation

At the trial runs, the limited keywords imposed restriction on searches. We had to apply expert knowledge to define useful keywords manually to increase the system performance. The raise of the statistic threshold may reduce the labour time more efficiently but also may lose some important information. When the number of cases is increasing, this restriction will be more evident. Currently, the queries in a specific building type are more predictable with the accumulation of keywords.

Moreover, all the semantic relations among cases or keywords need the pre-process due to the technical limitations, and then the system cannot directly respond the change of contents when a new cases or a new recommendation is appended into the system. It will be important to develop or apply more efficient clustering algorithm and interface for the online data mining for the further research.

4 DISCUSSION

The implementation of the semantic query mechanism of a case library described above provides the ground for the following discussion.
4.1 Explicit vs. Implicit Information

Keywords extracted from cases are still explicit information in some degrees, although it is not as clearly as the feature-pair data in general information, and inevitably may be short of the necessary information that users may be interested. To overcome this disadvantage, the expert domain knowledge can be applied to expand the completeness of keywords. Furthermore, the efficiency of keywords can be improved by collecting those terms in users queries to understand their interests.

To discover the implicit information of cases still must rely the basic relation of explicit information such as which keyword appears at which case and its contexts. An explicit keyword inputted by domain experts without its contexts perhaps can help users to retrieve a case and can not help them to understand this case. The FBS model is useful in establishing the relations among existing keywords and connecting them with the ontology.

4.2 Associative Reasoning

A case library customizes the artificial memory of cases for various purposes, and certainly limits the basic associative function of human natural memory. However, we do not attempt to make the association for users but help user to retrieve cases with their fragments of memory. For example, a user may forget a case named “Villa Savoye”. But he may have some memory fragments such as “a white house standing on a grass hill has a ramp and a roof garden in its interior”, and put this fragment into CBA system and retrieve some cases based on the keywords such as “white”, “grass hill”, “ramp”, “roof garden” extracted by system.

The catalyst of retrieving cases from memory is the association between the problems and the solutions in memory. The CBA system does not intent to provide the direct answers, but retrieves relevant cases to provide enough clues to inspire users to make further retrieving and reasoning. For example, a user looking for “Villa Savoye” may actually want to find cases in a suburban site. The CBA system can retrieves not only “Villa Savoye” but also similar one such as “Smith House”. Then, based on the contexts of keyword “grassy hill”, user can find relevant solutions about the view of living room and the placement of entrance. Similarly, user can find two different usages of “ramp” as interior passage and an entrance. This information can stimulate users to make their own associative reasoning to define design problems such as the relation among living room, entrance and the landscape of site. Users can also learn some knowledge such as the usage of a ramp.

4.3 Necessary Number of Cases in a Case Library

Currently, the case library in CBA contains 30-50 cases for each building type. For the learning purpose, the number of cases may not be the critical in a case library. But an appropriate number of cases or key cases will be necessary for referential purposes. Meanwhile, the number of keywords will not linearly increase with
number of cases, for example there are average about 30 keywords for a low-rise house case in CBA, but only about 75 distinct keywords for three cases. This phenomenon reveal the number of cases is not the crucial factor for the representation of a special type of cases. For the purpose of learning, the distribution of typical cases contains critical problems and solutions should be more important than the number of similar cases. Further investigation will be required.

The case library can be expanded to accommodate more similar cases for referential purposes. However architectural design generally involved in many different domain knowledge, a case library actually needs various subsets of cases in accordance with various building types. The increase of distinct building types should increase distinct keywords that can present the distinction among domain knowledge, and all cases should find some common keywords based on the same domain knowledge.

5 CONCLUSION

This study implements an idea of retrieving cases based on the semantic relationship of cases and investigate its feasibility for improving the keyword-based searching modality. Based on the prior discussions, the conclusions are as follows.

5.1 Implications for Building A Case Library

What should be included in the case library? The question addresses two issues of how to build a design case library: (1) the number of cases, or (2) the distribution of cases. The different types of cases involve different domain knowledge. For effectively establishing the cases library and representation the knowledge in cases, we should select the typical cases containing the critical information of special types. Observing the increases and repeats of extracting keywords among case can help us to determine which cases should be the more typical cases in the special type. The typical cases should have most distinct keywords co-occurring in most of cases of the special type.

Since the clustering of case can be recognized as the different classes of design problems or solutions of a special case type, then the number of cases clustering in a special type can help us to determine the number whether enough or not. Once the number of case clustering does not increase apparently with the increase of cases anymore, then the number of case should be enough for this case type.

5.2 Generality of Keywords

The contents of cases consist of three levels of information enable extracting implicit information from cases by the data mining technique. Then we can improve the efficiency of the keyword-based searching modality. The semantic relation of keywords in CBA system are not only play the critical role of cases retrieval, but also index the relevant contexts in cases for users to formalize their design problems.
and help them to understand the solutions in the cases. Namely users can associate
the contents of cases and their queries to make further reasoning. Furthermore, the
keywords can provide users as the entrance and the label of the implicit information
to help users to learn and memory the design knowledge in the cases.

5.3 Smart Query vs. Smart Answer

The smart answer often comes from the smart query (Riesbeck 1996). We do not
attempt to make a smart machine that can automatically reason about design itself,
or understand and answer various questions. Our major task is to improve the
functions of CBA to help users to retrieve most appropriate cases based on their uses
by the prior ranking of information in the system. Because the queries and results are
clustered and accumulated in accordance with the FBS model, the search mechanism
becomes smarter. Once users put their queries into the CBA system, the system
extract the keywords from queries and provide more relevant cases based on the best
guess of system by means of the semantic relation of keywords.

While we can map out the basic relations of the cases library, we still need the key
to access the library for exploring the essence of design information in the case
library. Therefore we will collect typical problem-solution scenarios based on the
contextual terms of keywords as the impetus for users to access different types of
relative cases in the future research in the future studies. Meanwhile, we will
incorporate different expertise for the online data mining when different building
type cases are collected.

REFERENCE:

York: Addison Wesley

for Semantic Retrieval of Web Data. Data & Knowledge Engineering, 28, 63-
84. Elsevier.

Problems in CAAD: Data Mining in i-CAADRIA. Proceedings of CAADRIA
2002, eds. A. Eshaq, W.K. Chee, K.T.K. Neo, M. Neo, S.N.S. Ahamad, 31-
38. Selangor, Malaysia: Prentice Hall.

Database in Architecture, eds. Alexander Koutamanis and Harry

