INFORMATION MINING WITHIN A CASE LIBRARY

Visual links of Correlation among Cases

CHIEH-JEN LIN[1], MAO-LIN CHIU[2]
[1]*Department of Interior Design, Tainan Woman's College of Arts & Technology, Yung-Kang City, Tainan, Taiwan*
jackyljz@seed.net.tw
[2]*Department of Architecture, National Cheng Kung University, Tainan, Taiwan*
mc2p@mail.ncku.edu.tw

Abstract. This paper is aimed to establish the visual links of correlations among design cases in a case library for architecture, CBA. The study found that the keywords extracted from cases usually present certain information related to design concept, knowledge, problems or situations; and therefore, the links of those keywords can represent the correlations of those concepts or knowledge. Then the links of keywords can help users to understand the correlation between those concept or knowledge and further to prompt the correlation of cases where contain the design concepts or knowledge. Based on the previous works, the collected cases are clustered by the semantic relationships of keywords extracted from cases and the links of cases are presented with the links of keywords. Furthermore, the links of those keywords and the ranking of those linkages can be visualized to represent the correlation among cases for helping users to retrieve appropriate cases and facilitate associative reasoning based on the information embedded in those cases. The interface implementation and feedbacks are discussed.

1. Introduction

1.1 EXPLICIT AND IMPLICIT INFORMATION IN CASE LIBRARIES

The existing case libraries, the information base of case-based systems, generally focuses on the collection of cases and the information retrieval functions for helping users to access potentially useful cases. However, what the key information is for user and where users can retrieve the useful information are usually less concerned in the development of case-based systems. Moreover, most of the information in system is explicit while there are more implicit and fuzzy information in cases, especially within the correlation among cases. The explicit information of case such as the site area, building scales, structural systems, materials, and building styles, can help users to determine whether the case is appropriate to apply or not, but usually is less useful for how to making design reasoning and cases adaptation. The implicit information, such as which cases are applying similar design techniques or concept, and which cases have the solutions of similar design problems or situations, are more contributive to inspire users to learn those techniques or concept and then to make their own reasoning or solutions. Therefore, how to extract this implicit information from cases and to present this information to users shall be more important than the collection of explicit information in the case-base design and reasoning. This paper is aimed to establish the visual links of correlations among cases in the CBA.

1.1 DESIGN INFORMATION IN CBA

This research is on the foundation of a case library, "Case Base for Architecture" (CBA) of office building and single house cases implemented on the web, as a case repository and a learning environment for case-based design. Based on the depth of knowledge, there are three levels of design information in CBA, namely "general", "analytic" and "recommendation", Figure 1.

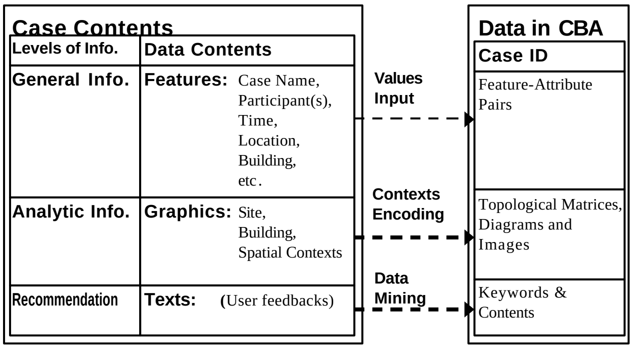


Figure 1. Abstracting of Case Contents in CBA

The general information of cases is the first and shallowest level of design knowledge in CBA that includes the most explicit attributes of case features, which are represented in the "attribute-value pairs" format in the CBA system similar as other case libraries.

The analytic information of cases is the deeper level of design knowledge in CBA that usually involves many non-text media materials, such as images, drawings and diagrams, which are difficult to directly convert into an indexical and searchable format. Therefore we will alternatively encode the information by attaching annotations, such as the structural prototype or other graphic diagrams, to represent the features of those non-text materials.

The "recommendation" of cases contains more conceptual and deeper degrees of design knowledge that is the user feedbacks or commentaries of cases from case creators, the professional critics or the users of the system. 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 KEYWORDS EXTRACTED FROM A CASE LIBRARY

For extracting the implicit information, we apply the data mining technique to extract a list of keywords of design concepts and knowledge from the text contents of the recommendation and analytic information of CBA. By clustering and ranking the semantic relations of those keywords on the strength of machine and addressing them reverse to selected cases, we reorganize the semantic relations among cases. (Lin and Chiu, 2003) Based on the results, we have built a primary query interface, and now we try to visualize the correlations among cases such as the sharing concept and techniques among cases and the correlations of design concept.

2. Visualization of the Correlations among Cases

2.1. THE SEMANTIC RELATIONS AMONG CASES

The conventional approach of information retrieval considers a document as a bag of words and ignores the semantic relations among them because of the difficulty of the natural language processing. The speech tagger algorithms, such as Brill's famous rule-based tagger (Brill 1992), can automatically tag the syntactic function of words in sentences and help the selection of feature words based on those tags, because we usually consider the nouns is more important than other syntactic functions of words in the representation of a documents. (Baeza-Yates, 1999) However, the tags of syntactic function and the syntactic analyses usually cannot help to understand the semantic relations of them.

The idea about the semantic relations among case is based on those keywords extracted from the recommendation and analytic information of cases. Since those keywords co-occur in a same sentence usually have semantic association on the description of a concept, then we can rank this semantic association ($s_{i,j}$) of two keywords (k_i, k_j) based on their co-occurring frequencies ($f_{i,l}$, $f_{j,l}$) in a case (c_l).in Equation (1).

$$s_{i,j} = \frac{\sum_{l=1}^{n} f_{i,l} \times f_{j,l}}{\sum_{l=1}^{n}((f_{i,l}^{2} + f_{j,l}^{2} - (f_{i,l} \times f_{j,l})) \times r_l(k_i, k_j))} \tag{1}$$

However, keywords co-occurred in a same sentence should have more semantic association with the description of a concept than co-occurred in a same paragraph but different sentences, so we apply the distance function $r_l(k_i, k_j)$ to distinguish them. If two keywords co-occurred in a same sentence, then $r_l(k_i, k_j)$ should be one. Furthermore, $r_l(k_i, k_j)$ should plus one for every co-occurrence in a same paragraph but different sentences, and $r_l(k_i, k_j)$ should plus two for every co-occurrence in different paragraphs of same information content.

2.2. BASIC SCHEMA OF DESIGN INFORMATION

In general, users may request different information at different design stages for various needs. The FBS (function, behavior, structure) schema is adopted in the definition of case content. (Maher et al., 1995) Because the collected cases in CBA, currently single low-rise houses, have similar functional requirements, the basic design problems or situations of architectural design can be summarized into three levels based on their hierarchies:

1. *Site Contexts*: The environmental factors around the building site are the largest scale and the first level of design problems. The conditions of the site contexts usually are the principal keys impacting or constraining the design results.
2. *Building Contexts*: The inner factors of the building site and the interactive factors among buildings and the site are the middle scale and the second level of design problems. The conditions of the building contexts usually are the secondary keys
3. *Spatial Contexts*: The arrangement, structure, construction and material factors of a building shaping spaces to resolute the design problems are the smallest scale and the third level of design problem. The conditions of the spatial contexts are the most detail but maybe the most concerned keys by the client and the designer impacting or constraining the design results.

Three primary keywords, namely “site”, “building” and “space”, are used to represent three levels of design problems or situations of case as the basis for searching and representation of design information. Then we can connect three primary keywords with their closest keywords based on their semantic association with Equation (1). Through the second connected keywords, we can connect them to the closest design cases or more associative keywords further to find more relevant information or cases, and then those selected keywords may become the representation of user's searching intentions. On the other hand, those sharing keywords among cases can be the representation of correlation among cases.

However, the information of cases may not contain any one of three primary keywords and lose the primary connection. To solve this problem, we apply two strategies to expand the primary connection of keywords: (1) manually assign keywords without any connection to one of three primary keywords by expert's domain knowledge or user's judgment; (2) apply the clustering algorithm to rank the semantic similarity between them and three primary keywords to cluster them into one of three primary classes in Equation (2).

$$sim(k_i,k_j)=\frac{\vec{v}_i \bullet \vec{v}_j}{|\vec{v}_i|\times|\vec{v}_j|}=\frac{\sum_{c=1}^{t} w_{i,c}\times w_{j,c}}{\sqrt{\sum_{i=1}^{t} w_{i,c}^2}\times\sqrt{\sum_{i=1}^{t} w_{j,c}^2}} \qquad (2)$$

Since the similarly semantic words should have the similar contexts, then we can compare the contextual words of keywords to find which keyword may have the most similar context with three primary keywords. Therefore we can gather the statistics of contextual words of keywords and convert the raw frequencies of them into the *tf-idf* weight, and further arrange those weights into a context vector $\vec{v}_i$. The similarity *sim(k$_i$,k$_j$)* of two keywords k_i, k_j can be quantified by the *cosine* of the angle between two context vectors.

2.3. THE METAPHORS OF VISUALUZATION

The goals of our visual interface are (1) to help users to make queries easier and (2) to inspire users to select appropriate cases faster and to make associative reasoning more effectively. On the one hand, we attempt to present and to reorganize the keywords extracted from cases in order to help users to present their searching intentions instead of inputting queries directly by users. On the other hand, we try to represent the relevant keywords among cases to help users to understand the sharing concepts among retrieved cases to determine appropriate cases and further to make their associative reasoning.

2.3.1 Query Space vs. Result Space

When searching in case library, there are always two sets of data: queries inputted by users to present their searching intentions and result cases retrieved from system to respond user's queries. Therefore we chose to horizontally divide the display space of our visual interface into two areas: left side is the "Query Space" that users can select keywords and follow their semantic association to represent their searching intentions, and right side is the “Result Space” that system display relevant cases retrieved by keywords in the “Query Space” based on their semantic association and keyword’s weighting.

2.3.2. Visualization of Keyword’s Semantic Association in Query Space

In the “Query Space” each keyword is displayed as a rectangle label, and three primary keywords are always laid vertically at most left side, namely “site”, “building” and “space” with this order, and play the role of starting searching points and the essential clues for learning and understanding the concepts of cases especially useful for beginners. From these keywords, users cannot only find relevant cases, but also can further to finds more relevant keywords based on their semantic associations.

For assisting user to select appropriate keywords easier, we chose to attach each keyword, expect three primary keywords, with a linear measure below it to indicate the number of their relevant keywords. Moreover each keyword is connected to relevant keywords and cases with a gray line that their width indicates the ranking of its semantic associative relation with other keywords and cases. The wider connecting line presents the more ranking of the semantic associations among keywords or cases.

Figure 2 demonstrates that two selected keywords from primary keywords, such as “slope” and “lake” from “site” to present a “site context” that a hillside slopes down to a lake, and retrieve two cases, such as Case A “Douglas House” by Richard Meier and Case B “Residence in Riva San Vitale” by Mario Botta.

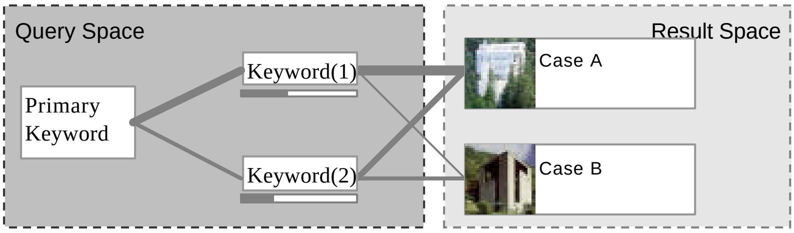


Figure 2. Representations of Selected Keywords in “Query Space” and Retrieved Cases in “Result Space”

2.3.3. Visualization of Implicit Correlations among Cases in Result Space

If a keyword excluding three primary keywords is selected in the “Query Space” then the relevant case will be displayed at the most left side of the “Result Space” based on the *tf-idf* weighting of all selected keywords in cases. More keywords are selected then less relevant cases are retrieved. Those initial cases are not only the retrieved results by keywords in “Query

Space", but also can play the role of starting point for searching more relevant cases. From these cases user can further to find more relevant cases based on sharing keywords among cases.

For assisting user to select appropriate cases more effectively, we chose to present each case as a small icon with its most representational picture and to attach case name with it. Furthermore we connect the relevant cases in "Result Space" with a gray thin line and attach their sharing keyword that are relevant to those selected keywords but not displayed in "Query Space" to present the implicit semantic relations among retrieved cases. Similarly the wider connecting line presents the more ranking of the semantic associations among keywords or cases. Figure 3 demonstrates three sharing keywords, such as "entrance", "bridge" and "view" etc., represents the implicit semantic relations between two retrieved cases, such as "Douglas House" and "Residence in Riva San Vitale".

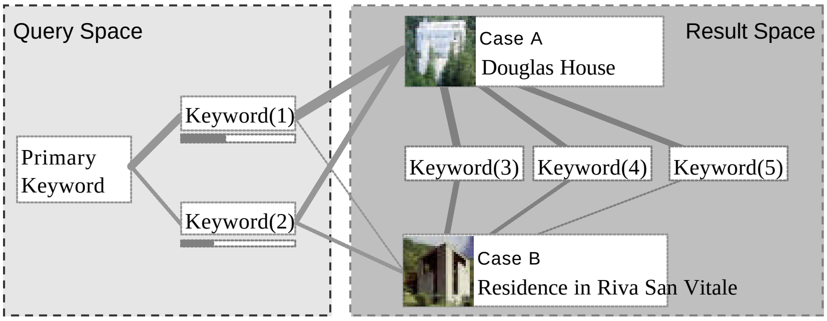


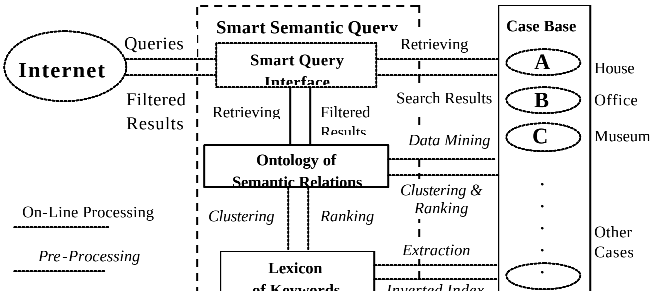
Figure 3. Implicit Semantic Relations among Retrieved Cases in "Result Space"

Through trailing the semantic associations of keywords in "Query Space", users can present their searching intentions step by step and find most interesting cases in system. Through trailing those sharing but implicit keywords among cases in "Result Space", user can more effectively recognize implicit sharing concepts among cases and further discover useful information in cases more conveniently.

3. System Implementation

3.1. FRAMEWORK OF CBA

The framework of CBA system is illustrated in Figure 4. After the pre-processing of text context of "analytic" and "recommendation" information of cases in CBA, 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 associative relations of keywords. Then our visual interface provides for the query interface and the representation of result of retrieved cases. By applying the techniques of graphic interface based on Macromedia's Flash to integrate with the Microsoft Access database through the techniques of JAVA and ASP, we implement the visual interface of CBA on the web.

Figure 4. Framework of CBA.

3.2 QUERY INTERFACE AND PATTERN RETRIEVAL

Based on the ontology of semantic associative relations of keywords, the CBA system currently implements three levels of search functions such as (1) general search by feature matching and full-text keywords matching; (2) the keywords browser that allows users freely surfing the whole keywords list and the ranking of semantic associations of keywords to retrieve any cases that users are interested; (3) a visual, smart and interactive query interface for user to present their searching intentions.

The "Query Space" of our visual interface cannot only provide for a visual query interface, but also provide an interactive representation for user to discover the potential and relevant situations and issues in cases that is especially useful for beginners in learning and understanding the design problems of cases. The "Result Space" of our visual interface can reveal the potential clues of semantic relations among case to help users to recognize implicit and potential patterns in retrieved cases that are more useful for general designers and experts in searching for useful solutions and design knowledge in retrieved cases. Then users can follow those clues to explore among relevant cases and inspire them to make more associative reasoning.

Figure 5 demonstrates the query interface of CBA: the keywords browser (left) and visual query interface (right). The keywords browser is a multi-columns table that reveals the relevant keywords and the ranking of semantic associative intensity in order to simplify the complex network representation.

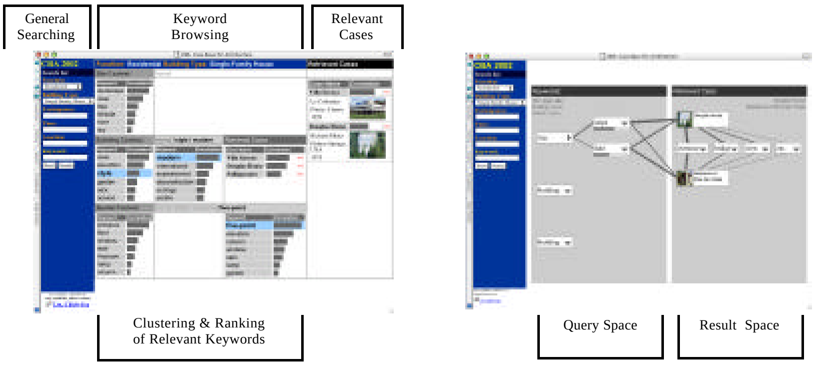


Figure 5. CBA Interface: The Keywords Browser (Left) and Visual Query Interface (Right).

3.3 PRELIMINARY EVALUATION

Although some attempts have been made in CBA, there are some problems in our approach: (1) the automatic filtration of noise keywords by machine is restricted; and (2) the interpretation the meanings and contexts of keywords in cases is required. Therefore, the keyword approach requires some expert knowledge for enhancing the correlation among keywords.

The noise keywords, not the “stopword” such as a preposition or a determiner, may have some helping explanatory functions for useful keywords, design problems and situations, but are not easy to directly recognize their significance in design problems or situations and become a noise in early searching stage. On the other hand, some useful keywords may be meaningful for general designers or experts but not easy to be immediately understood by beginners. Therefore between keywords and information in cases there need some explanatory data, namely “meta keywords” of keywords such as three primary keywords, to help user to recognize the meaning of relevant keywords in cases.

4. Discussion

The implementation of our visual query interface of a case library described above provides the ground for the following discussion.

4.1. META KEWORDS OF IMPLICIT IMFORMATION

Keywords extracted from cases, however, are still explicit information in some degrees, although it is not as evident as the feature-pair data in general information, and inevitably may be short of the necessary clue for users to recognize interesting information or understand their significance in cases. For reducing this disadvantage, the expert domain knowledge should be applied to establish the more explanatory and indicative “meta keywords”, like the thee primary keywords, to make keywords extracted from case more comprehensible especially for beginners.

However, the “meta keywords” inputted by experts would be difficult to automatically establish their semantic connections with those keywords extracted from cases because there are no direct links among them stored in the system. The traditional method to make this connection would rely on experts to review all keywords and their information in cases to categorize them manually. Our visual interface can make this task easier and more effectively because we can just review those keywords and their semantic associative relations among retrieved cases to categorize them.

4.2. SEMANTIC ASSOCIATION IN CASE LIBRARY

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 only attempt to establish a smart and interactive interface to assist user to retrieve interesting cases, but also endeavor to provide semantic clues for user to inspire them to make their own association.

The catalyst of case-based design is the association between the design problems and situations occurring to designers, such as the similar situations of site, “a hillside slopes down to a lake” for example, with the solutions in cases, such as “a red bridge connect to upper story as a entrance” (“Residence in Riva San Vitale”) or “living room faces the lake” (“Douglas House”) for example. Through following the semantic associative connections among keywords, user can easier and more naturally discover the association of them to make their own reasoning.

4.3. DESIGN INFORMATION IN A CASE LIBRARY

Two major issues about the implications of case library are (1) the number of cases, or (2) the distribution of cases. The case library can be expanded to accommodate more similar cases for referential purposes. However the number of keywords and their semantic associative connections would not linearly increase with the number of cases. Therefore the number of meaningful semantic associative connections among cases should be more important than the number of cases for providing useful design information.

More meaningful semantic association among keywords and cases mean that there are more potential information stored in system. Therefore the most contributory cases should be those case that have most number of meaning keywords and semantic association among them with other cases. Our visual interface can reveal the contribution of semantic association of a case and then help us to select appropriate paradigmatic cases.

5. Conclusion

In conclusion, this study implements the visual query interface based on the semantic associative relationship of cases and investigate its feasibility for improving information mining in case library. Based on the prior discussions, we have the following findings.

5.1 IMPORTANCE OF VISUAL LINKS

The visual links in our visual interface presents the semantic associations among keywords and cases that can help user to easer and more effectively recognize what and where is the important and interesting information in case library. More dense connecting lines among keywords and cases indicate the more important and interesting keywords or cases that the former presents the brief concepts what are the contents of information, and the latter indicates the locations where are information embedded.

5.2. NECESSARY OF META KEYWORDS

Essentially those keywords in shallower levels of our visual interface should have more distinct indication and those keywords in deeper level should have more definite interpretation. However, the complete explanatory hierarchy of keywords still cannot establish by our approach. Therefore most keywords in CBA still need the explanatory and indicative “meta keywords” to help user to immediately recognize potential useful keywords and expedite the user’s searching and inference. The visual interface can make this task easy but it still needs more investigation.

5.3. ASSISTANCE IN ASSOCIATIVE REASONING

Through tracing the semantic associative connections among keywords and cases, general designers and experts can associate their queries with the contents of cases and to make further reasoning, and beginners can retrieve the relevant contexts in cases to formalize their own design problems and help them to understand the solutions in the cases. Therefore our visual interface plays a role of an assistant in user’s associative reasoning rather than the inferential agent of traditional Artificial Intelligence.

5.4. USEFUL INFORMATION DISCOVERY IN CASE LIBRARIES

Whether the design information embedded in design cases is useful or not depends on the user’s intentions and design situations encountered. However, user’s intentions would change and the design problems would be redefined in different design stages, then the useful information would vary too, especially for beginners that they are lack for necessary design knowledge to understand design situations and formulate their own design problems. Through exploring in the visual interface, users have chances to explore their different intentions by review retrieved relevant keywords and cases and further to modify and ascertain their intentions to determine whether those information is useful.

5.4. FUTURE DEVELOPMENT OF CBA

For improving the efficiency of our visual interface in discovery useful information in case library, the establishment of "meta keywords" lexicon, that cant help us to categorize keywords in CBA to develop navigational assistant for large collections of case, should be the next step of development. Therefore we will collect typical problem-solution scenarios to find the typical requirement of design information and undertake to build the "meta keywords" lexicon and their semantic relations among cases.

References

Lin, C.J. and Chiu, M.L.: 2003, Smart Semantic Query of Design Information in A Case Library, *Proceedings of CAAD Futures 2003,* Tainan, Taiwan. (Accepted)

Chiu, M.L., C.J. Lin, T.S. Jeng and C.H. Lee.: 2002, Re-Searching the Research Problems in CAAD: Data Mining in i-CAADRIA. *Proceedings of CAADRIA 2002,* Prentice Hall, Kuala Lumpur, Malaysia, pp. 31-38.

Chiang, R.H.L, C. E. H. Chua, and V.C. Storey.: 2001 A Smart Web Query Method for Semantic Retrieval of Web Data, *Data & Knowledge Engineering* 28, Elsevier, pp. 63-84.

Baeza-Yates, R. and B. Ribeiro-Neto.: 1999, *Modern Information Retrieval,* Addison Wesley, New York.

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

Riesbeck, C. K.: 1996, What Next? The Future of Case-Based Reasoning in Post-Modern AI. David B. Leake (ed), *Case-Based Reasoning: Experiences, Lessons, & Future Directions,* the MIT Press, Cambridge, pp. 371-388.

Maher, M. L., Balachandran, M.B., Zhang, D.M.: 1995, *Case-based Reasoning in Design,* Lawrence Erlbaum Associates, Inc., New Jersey.

Brill, E.: 1992, A Simple Rule Based Part of Speech Tagger, in Proceedings of the Third Conference on Applied Natural Language Processing, ACL, Trento, Italy.