Evaluating CBR Systems Using Different Data Sources: A Case Study

Mingyang Gu and Agnar Aamodt

Department of Computer and Information Science, Norwegian University of Science and Technology, Sem Saelands vei 7-9, N-7491, Trondheim, Norway
{mingyang,agnar}@idi.ntnu.no

Abstract. The complexity and high construction cost of case bases make it very difficult, if not impossible, to evaluate a CBR system, especially a knowledge-intensive CBR system, using statistical evaluation methods on many case bases. In this paper, we propose an evaluation strategy, which uses both many simple case bases and a few complex case bases to evaluate a CBR system, and show how this strategy may satisfy different evaluation goals. The identified evaluation goals are classified into two categories: domain-independent and domain-dependent. For the evaluation goals in the first category, we apply the statistical evaluation method using many simple case bases (for example, UCI data sets); for evaluation goals in the second category, we apply different, relatively weak, evaluation methods on a few complex domain-specific case bases. We apply this combined evaluation strategy to evaluate our knowledge-intensive conversational CBR method as a case study.

1 Introduction

As summarized in [1], AI research is an empirical process: selecting a task incorporating intelligence features, building a system exhibiting these features, and evaluating the system in different task environments. After an intelligent system is constructed, it is necessary to evaluate whether it does what we expect it to do and how good its performances is. Cohen and Howe [2] extend the importance of evaluation from assessing the performance to guiding the different AI research phases.

Evaluation methods for intelligent systems include statistical evaluation (inductive evaluation), theoretical analysis, ablation evaluation, tuning evaluation, limitation evaluation, direct expert evaluation and characteristic analysis [3–5]. The ideal evaluation method among them is statistical evaluation; that is, to execute the constructed system in different task environments in order to investigate its performance in different application contexts. This method is difficult to apply, in general, to case-based reasoning (CBR) [6, 7] because of the typical complexity of CBR applications. The complexity comes from two aspects [8]: the CBR system itself is complex, and the task domain where it operates is also typically complex and ill-structured. The complexity of the application domain makes it difficult and expensive to construct a case base, especially for knowledge-intensive CBR systems [9] that demand a significant knowledge engineering effort. Because of the complexity and heterogeneity of CBR systems, transplanting a case base from one CBR system to another also needs considerable adaptation work. Therefore, it is very hard to construct or transplant a number of complex case bases to use in a statistical evaluation. For these reasons, the evaluation of a CBR system is, to a large extent, based on one or a few case bases, which can only provide limited evidence.

Aha [10] provides a method to generalize the evaluation result of an AI system, which is based on one (or a few) data sets. In this method, a set of dimensions are identified to describe the original data set, and a data set generator is created to produce many artificial data sets with predefined values on the identified dimensions. The target system is executed on the artificial data sets, and its performances are recorded. The relations between differences of the system performance and changes of the dimension values are studied, and a set of rules are generated to describe the conditions under which the performance differences hold. Applying this method into CBR researches needs substantial efforts since it is difficult to artificially construct a set of complex case bases with the predefined dimension values.

When we look into the details of the evaluation process for CBR research, we find that there are usually multiple evaluation goals. For instance, this includes the efficiency of the similarity calculation method, the validity of the adaptation method, the problem solving efficiency on the target application domain, the usability or human friendliness, and the individual contributions of various system components. Further, different evaluation goals are related to different application scopes. Some goals are domain-dependent; that is, they need to be evaluated on the target specific application domain, for example to determine whether the general domain knowledge can improve, for instance, the similarity matching using a knowledge-intensive method [11], or make an explanation to the user more understandable [12]. Other goals are domain-independent, for instance whether the sustained learning process in CBR can improve problem solving capability. For the domain-independent goals, we can evaluate them on either complex case bases or simple case bases. There are plenty of such simple case bases, for instance, the data sets available at the UCI repository [13], and there are many examples of research contributions evaluated by this data sets within CBR community [14–17].

We propose an evaluation strategy for CBR research aiming to assess these two types of evaluation goals (domain-dependent and domain-independent) based on different data sources and using different evaluation methods. For the domain-independent evaluation goals, we use the statistical evaluation over many simple data sets, while domain-dependent goals are evaluated on one or a limited number of complex case bases using multiple weak evaluation methods. That is, this strategy combines a statistical evaluation method with many simple case bases, and alternatively combines limited number of complex case bases with multiple weak evaluation methods. This evaluation strategy can provide solid evidence for both the domain-independent goals and the domain-dependent goals. For the domain-independent goals, the evaluation power comes from the statistical justification. For the domain-dependent goals, the solidity comes from whether all the multiple weak evaluation methods can output positive outcomes.

As part of our current research, we have designed and implemented a knowledge-intensive conversational case-based reasoning (KI-CCBR) system which can capture and utilize general domain knowledge to support an efficient and natural conversation process to complete the case retrieval task. In this paper, we use our proposed evaluation strategy to evaluate this KI-CCBR method as a case study.

In the next section, we give a short introduction to the evaluation methods we have used. In Section 3, we briefly introduce our KI-CCBR method and identify the relevant evaluation goals. In Section 4, we report how we use 36 UCI data sets to show that the two domain-independent evaluation goals, lazy dialog learning and query-biased similarity calculation, can improve conversation efficiency of CCBR in general. We also evaluate the KI-CCBR method on a case base of image processing software components, within a system designed to support component reuse in software design. Three different evaluation methods are used: a characteristic analysis is used to see whether the system meets the requirements of a conversational diagnosis system; a direct domain expert assessment is used in order to see whether the KI-CCBR method can provide a natural conversation process; and a simulated ablation study is adopted to evaluate whether KI-CCBR can improve the conversation efficiency and how much each knowledge-intensive module contributes to the total improvement. We conclude in Section 6.

2 Introduction to the Evaluation Methods

The purpose of an evaluation process is to assess a system, with reference to some selected baseline, to see whether the performance of the system is accepted or improved. In this section, we introduce the evaluation methods used in our study.

2.1 Statistical Evaluation (Inductive Evaluation)

The basic statistical evaluation process is one in which we define one or more performance measures, execute both the new system and the baseline system on many different data sets, and calculate the percentage of the data sets on which the new system gives better performance, or test statistical significance in relation to predefined hypotheses. Statistical evaluation is a proper method to support the claim of generality of a system's benefits or advantages. This method is a strong evaluation method and is frequently used in many scientific disciplines such as psychology or biology. Cohen [4] gives detailed information about how to apply this evaluation method for AI research.

2.2 Characteristic Analysis

For a certain type of intelligent system, what characterizes it are usually discussed and gradually agreed upon by researchers in that field. Analyzing whether and to what degree a system can support the relevant characteristics is one approach to evaluate the system with respect to its possible performance. For example, if a CBR system support all the four 'RE-' phases [6], we may claim that this system is a full-cycle CBR system.

2.3 Direct Expert Evaluation

When a test system can produce more acceptable solutions than we can possibly generate beforehand as a baseline [3], or the evaluation measures involve human common sense or psychology aspects, one method is to invite domain experts to use the system and report back their subjective assessments. This is a weak evaluation method because of experts' overly generosity and their unrepresentativeness of typical users. One way to balance this shortcoming is to select experts using different criteria, or experts from different related domains.

2.4 Ablation (Lesion, Substitution) Evaluation

Ablation evaluation [2, 8] is a method for analyzing the contributions of different modules of a system to the total performance improvement. In this type of evaluation, one or more modules are de-activated, removed or replaced by other comparable modules to observe changes on system performance. This method was used to evaluate the PROTOS system [18] and the SIROCCO system [19]. One difficulty in applying this evaluation method is that it is not always feasible to remove or de-active particular modules in a system because of the interdependence among modules.

3 Knowledge-Intensive Conversational Case-Based Reasoning

3.1 Research Overview

Conversational case-based reasoning (CCBR) [20] is a special type of CBR, which emphasizes the difficulty to appropriately describe a new problem, i.e. to define a new case. CCBR alleviates it through providing a mixed-initiative interactive process to guide users to incrementally construct a new case description that is sufficient to complete the case retrieval task.

In CCBR, an initial new case (only one or few features) is specified and used to retrieve a set of most similar cases from the case base. A group of discriminative questions are identified based on the returned cases (transformed by the features that have values in the returned cases but not in the current new case), and ranked according to their capabilities to discriminate the stored cases. Both the returned cases, sorted according to their similarity values, and the ranked questions are displayed to the user. The user either finds a satisfactory stored case, which then terminates the case retrieval phase, or chooses a question to answer. The newly gained answer and the current new case are combined together to construct an updated new case. A new round of retrieval and question-answering is started, and this continues until the user finds a satisfactory stored case or there are no discriminative questions left for the user to choose.

A major research concern in CCBR is how to select the most discriminative questions [14, 21] and ask them in a natural way [20, 22, 5] to alleviate users' cognitive load demanded in the conversation process. Most of the methods used to select questions now are knowledge-poor (KP); that is, only statistical metrics are used. In our research, we study the possibility of using general domain knowledge in the conversation process [23]. We identify the following four tasks for which general domain knowledge can be used to improve the conversation process:

- **Feature Inferencing**: The features that can be inferred from the current new case description should be added into the new case description, instead of posting users questions.
- **Knowledge-Intensive Question Ranking**: The semantic relations among discriminative questions should be taken into account during question ranking. For instance, if one answer of question A, A_a, can be inferred out by one answer of question B, B_a, question B should be asked before question A.
- **Consistent Question Clustering**: The questions that are connected by some semantic relations, for example, a causal relation or subclass relation, should be grouped and displayed together, so that users can inspect them together and select which one to answer first.
- **Coherent Question Sequencing**: If a question from a higher level node in a taxonomic structure is asked in the current question answering cycle, the question one level lower should be asked in the next cycle, instead of inserting other unrelated questions between them.

We classify similarity calculation methods in CBR into three categories, according to the scope of the features that are taken into account during similarity calculation:

- **Query-Biased Similarity Methods**: Only the features appearing in the current new case (query) are taken into account during similarity calculation.
- **Case-Biased Similarity Methods**: Only the features appearing in the current stored case are considered during similarity calculation.
- **Equally-Biased Similarity Methods**: All the features appearing in both the current new case and the current stored case are taken into account during similarity calculation.

We emphasize the special characteristic of the new case, partially specified, in CCBR. If the features which have not yet been assigned values in the new case, are considered in the similarity calculation, the similarity method will be biased to those cases with fewer such features, instead of to those that most satisfy the current new case (users' attention focus). So in order to avoid the negative influence of these features, we argue that the query-biased similarity calculation method is more suitable for CCBR than the case-biased or equally-biased similarity calculation methods [24].

In addition, we introduce a lazy dialog learner into CCBR [25], which is capable of capturing and storing previous successful conversational case retrieval processes and reusing them in the later conversational case retrieval tasks.

The implemented KI-CCBR method has been recently tested in an image processing software component retrieval application [26].

3.2 Identified Evaluation Goals

As discussed in Section 1, the evaluation goals of this KI-CCBR method are classified into two categories. The first category contains the evaluation goals that are valid for CCBR research in general; that is, domain-independent: whether the query-biased similarity calculation method and the lazy dialog learner can improve the efficiency of CCBR. The second category includes the evaluation goals that rely on a specific application domain, the image processing component retrieval application. This includes whether the KI-CCBR method meets the requirements of a conversational diagnosis system, whether the KI-CCBR method can provide users with a natural question answering process, whether the KI-CCBR method can achieve higher efficiency compared to the knowledge-poor CCBR method, and how the different knowledge-intensive modules contribute to the total achievement.

In Section 4 and Section 5, we will report how we choose different evaluation methods and test case bases for the identified evaluation goals.

4 Statistical Evaluation on UCI Data Sets

In an attempt to evaluate whether the query-biased similarity calculation method and the lazy dialog learner can improve the efficiency of CCBR in general, we choose the statistical evaluation method to see whether these methods can achieve higher efficiency than their competitors on multiple simple case bases.

In order to evaluate which similarity calculation method (query-biased, case-biased, or equally-biased) is more suitable for CCBR, we implement three variants of CCBR within Weka [27], each of which uses a particular similarity calculation method. In order to evaluate whether the dialog learning mechanism can improve the conversation efficiency, we implement two more variants of CCBR also within Weka, one of them using our dialog learning mechanism and the other not. We summarize the statistical evaluation to these two topics in this paper, and more detailed information can be found in our earlier studies [24, 25].

The simple case bases we test are 36 classification data sets[1] provided by Weka, originally from the UCI repository [13]. Some of these case bases have been used to test conversational CBR methods in [14, 16, 17]. Aha, McSherry and Yang [28] argued that the typical case bases in CCBR applications are irreducible and heterogeneous. From our perspective, it is not necessary for case bases in CCBR to have these characteristics. For instance, in one typical CCBR application domain, fault diagnosis, it is natural for two types of faults to share the same solution, which means the case base is reducible. Heterogeneity is only the characteristic of one type of case bases in CCBR, which is the necessary condition to apply the occurrence-frequency metric [20] in discriminative question selection. However, the entropy based question selection methods, which are adopted by more CCBR researches [29, 21, 30], require all the cases having the same structure (homogeneous).

The human-computer conversation process is simulated using leave-one-out cross validation (LOOCV). LOOCV is an extreme variant of K-fold cross validation, which splits the entire n cases in one case base into n subsets, each containing only one case. In each evaluation cycle of LOOCV, the test case, q, is taken as a description of a new problem, referred to as the target case. Before the retrieval starts, a part of the problem description of q, a subset of the $< feature, value >$ pairs (10%), is taken out to construct an initial new case. This initial new case is used for retrieval from the test case base containing the remaining cases in the original case base. If the base case, with respect to the target case, is returned as the most similar case, or is in the returned most similar case group, the retrieval process is terminated successfully. Otherwise, the question generating and ranking module will output a set of sorted discriminative questions. A predefined question selection strategy is used to select a question from the discriminative question list, for example selecting the first question. The $< feature, value >$ pair corresponding to the selected question is chosen from the target case q and added into the current new case to form an updated new case. Based on the updated new case, a new round of retrieval is started. The retrieval, question selection and answering process will continue until the successful condition or failed condition (there are no $< feature, value >$ pairs left to choose) is met.

The average session number of the conversations simulated by the total set of cases in one case base is taken as the main criterion to assess the performance of a CCBR method on that case base [20, 14].

The successful termination condition of LOOCV is that the base case appears in the first returned case group (k cases). If the query biased similarity method is used, especially in the beginning phase of the retrieval process, the number, m, of the cases that exactly match the partially-specified new case (and are thus equally similar) may be larger than k. In this situation, the simulated process randomly returns k out of them. This setting may be arbitrary. Ferguson and Bridge [31] suggest a method to abandon exact similarities in favor of prefer-

[1] For the evaluation of the lazy dialog learner, we drop off the 4 biggest case bases simply because they need too much execution time.

ence relations between cases. In our case, the successful termination condition is acceptable since the final statistical result is computed from multiple cases and case bases using the same successful condition.

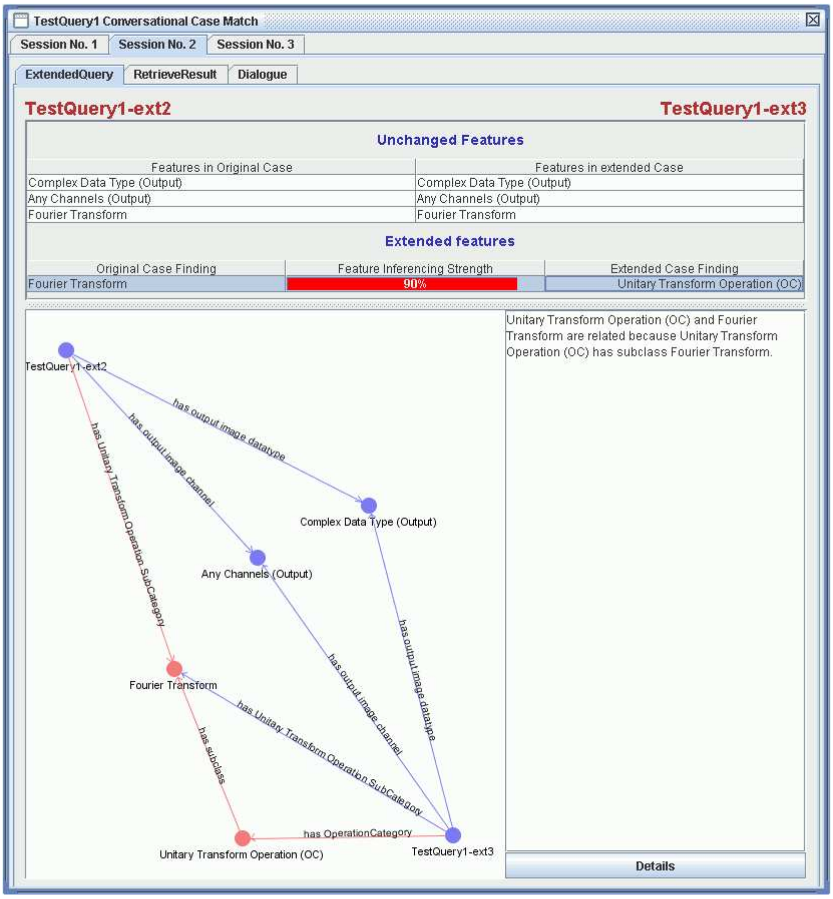
For the evaluation of the similarity calculation methods, in 31 out of total 36 case bases, the CCBR using query-biased similarity method gets better performance than the other two methods (case-biased and equally-biased similarity methods). For the assessment of the lazy dialog learner, in 29 out of total 32 case bases, the CCBR process with the lazy learner gets better performance than that without the learner. In this experiment, we execute the LOOCV two rounds with the aim to evaluate the ability of the lazy learner to learn in a long term. The results show that the lazy dialog learner is sustainable and the dialog case base is maintainable; that is, with the dialog learning going on, the dialog learner achieves better performance and the lazy learner requires fewer dialog cases to be stored in the dialog case base. For all the above comparisons, we have carried out the significance tests on the tested case bases (t-test with the significance level 0.01), and the results give us supportive evidence (all the observed differences in performance are significant).

5 Evaluating the KI-CCBR Method on an Image Processing Software Component Retrieval Application

We have implemented our KI-CCBR within the CREEK system [26][2]. We choose image processing software component retrieval, exemplified by retrieving components from the DynamicImager system [32], as the evaluation domain to assess the domain-dependent evaluation goals. DynamicImager is an image processing development and visualization environment, in which different image processing components can be combined in various ways. Currently, the components in the system are categorized according to their functions, and users select each component by exploring the category structure manually. A knowledge base has been constructed through combining image processing domain knowledge with 118 image processing components extracted from DynamicImager. In this knowledge base, there are 1170 concepts, 104 features and 913 semantic relationships, using approximately 20 relation types (e.g. has subclass, has part, causes).

As illustrated in Fig. 1, a conversational retrieval process contains one or several conversation sessions, and for each session, there are three window panes to move between in the computer interface. The ExtendedQuery pane is used to show how a new case is extended through feature inferencing, and to display a detailed explanation of why a new feature is added into the case. Based on the extended new case, a set of stored cases are retrieved and displayed in the RetrieveResult pane. In this pane the user can inspect the explanations about how the similarity values are computed. If a user is not satisfied with the retrieved cases, she can go to the Dialogue pane, where the discriminative questions are ranked using both the knowledge-intensive question ranking method and statistical metrics, and adjusted by the consistent question clustering and coherent

[2] The dialog learning mechanism is not implemented in our KI-CCBR.



TestQuery1-ext2

TestQuery1-ext3

Unchanged Features

Features in Original Case	Features in extended Case
Complex Data Type (Output)	Complex Data Type (Output)
Any Channels (Output)	Any Channels (Output)
Fourier Transform	Fourier Transform

Extended features

Original Case Finding	Feature Inferencing Strength	Extended Case Finding
Fourier Transform	90%	Unitary Transform Operation (OC)

Unitary Transform Operation (OC) and Fourier Transform are related because Unitary Transform Operation (OC) has subclass Fourier Transform.

TestQuery1-ext2
has output image datatype
has output image channel
has Unitary Transform Operation SubCategory
Complex Data Type (Output)
Any Channels (Output)
Fourier Transform
has subclass
has output image channel
has output image datatype
has Unitary Transform Operation SubCategory
has OperationCategory
Unitary Transform Operation (OC)
TestQuery1-ext3

Fig. 1. The conversational case retrieval process in our KI-CCBR

question sequencing processes. After the user selects a discriminative question and submits her answer, a new conversation session is started based on the updated new case by combining the newly gained answer with the previous new case.

5.1 Characteristic Analysis as a Sequential Diagnosis System

The CCBR process is basically a sequential diagnosis process: as more and more problem features (evidence) are identified and added into the new case, the system can identify the correct diagnosis (the base case) with more confidence. McSherry [5] identifies seven desirable features (or characteristics) of an intelligent sequential diagnosis system. Our implemented system supports all of these characteristics.

- **Mixed-Initiative Dialogue**: Users, particularly professional users, are unlikely to accept a conversation partner (or intelligent system) who keeps asking a lot of questions. Instead, users prefer playing a positive role in the conversation, i.e. to control the conversation or to volunteer data at any stage of the conversation. Permitting users to select a question to answer from a list is a form of mixed-initiative dialogue which is supported by most CCBR applications. In addition, our method allows users to cancel or modify their answered questions (go to the specific session window and remove or reselect the answered entities). Furthermore, our method clusters related questions together, so that users can inspect the questions with different difficulty levels, and select one to answer according to their expertise levels.
- **Feedback on the Impact of Reported Evidence**: It is unacceptable if users get no feedback from an expert (or intelligent system) after they provide more evidence. In our method, after the user answers a question, or modifies the initial new case or previous answered questions, the case retrieval process and the question ranking, sequencing, clustering processes will run immediately. The returned cases and discriminative questions will be based on the updated evidence.
- **Relevant Dialogue**: The questions asked by an intelligent partner should be relevant to the context of the problem provided by the user. We assume that only the features appearing in the most similar cases are relevant. Therefore, our method generates discriminative questions based only on the most similar cases, instead of all the cases in the case base.
- **Consistent Dialogue**: The questions that can be answered implicitly by the current partially specified new case should not be prompted again. Otherwise, the conversation efficiency is reduced, and users are unlikely to trust a conversation partner that repeats previously implicitly answered questions. Furthermore, if a user provides an answer to a question that is not consistent with that inferred from the current new case, the content of the new case is not consistent any more. The feature inferencing process in our method guarantees that this type of dialog inconsistency will not occur, by ensuring that these types of questions will not be asked.
- **Explanation of Reasoning**: In order to improve users' confidence in the results of an intelligent system, it is important to provide an explanation of how results are derived [33, 34, 5, 12]. Our KI-CCBR method provides the following explanations: why a new feature is added into the current new case description through feature inferencing, why two different feature values are partially matched through knowledge-intensive case matching [11], why a question is ranked with highest priority in the coherent question sequencing, and why two questions are grouped and displayed together through consistent question clustering.
- **Tolerance of Missing Data**: Missing data stem from two aspects. First, the cases in the case base may contain missing features. Our system's partial matching process can tolerate this type of missing data. We adopt the occurrence-frequency metric [20] as the knowledge-poor question ranking method, which basically takes the advantage of the presence of missing fea-

tures. In addition, our explanation-driven reasoning process [23] exploits general domain knowledge, which may itself be incomplete. Another source of missing data is the user's incapability to answer every question due to the unavailability of some observations, the user's lack of expertise, or need for an expensive test to obtain the answer. Our method tolerates this type of missing data through permitting the user to choose candidate questions to answer, instead of forcing her to answer them in a fixed sequence.

- **Sensitivity Analysis**: The uncertainty that is inherent in the dialogue process, as well as the possible uncertainty in the user's answers to questions, means that support for sensitivity analysis is essential. Our method supports sensitivity analysis through allowing users to modify previously specified features (answered questions) and re-execute the retrieval and question-answering process in order to inspect the possible influences of the updated information.

5.2 Domain Expert Evaluation of the Psychological Goals

Evaluation Goals of KI-CCBR related to psychology include the user's cognitive load, the 'natural' question-answering process, and the user's confidence in the final results. We adopt a relatively simple or weak evaluation method, a so-called direct expert evaluation [2], for these evaluation goals.

We invited two experts from the software engineering domain, and two experts from the image processing domain, to test our system. Given a set of image processing tasks, these domain experts were asked to retrieve image processing components using both a one-shot CBR-based retrieval method and the multiple shots KI-CCBR method. After doing so, they were required to fill in a form to describe their subjective evaluation of the implemented system[3]. The resulting analysis of the collected feedback forms suggests that:

- Based on the same initial new case, the KI-CCBR method can achieve more useful results;
- The reasoning transparency provided by the explanation mechanisms in KI-CCBR improves user confidence in the retrieved results;
- The feature inferencing, consistent question clustering and coherent question sequencing mechanisms provide users with a natural question-answering process, which helps to alleviate their cognitive loads in retrieving components interactively;
- The straightforward question-answering query construction process helps to reduce users' cognitive load in constructing queries, thus enabling users with limited domain knowledge to retrieve suitable components.

[3] The hypotheses list and the feedback form can be found at http://www.idi.ntnu.no/∼mingyang/research/CCRM_Evaluation.pdf

5.3 Ablation Evaluation Using Leave-One-In Cross Validation

In order to show that the KI-CCBR method does improve the conversation efficiency by reducing the length of conversation sessions compared to knowledge-poor CCBR, we execute another cross validation on the image processing component retrieval application. Unlike the LOOCV we introduced in Session 4, we adopt leave-one-in cross validation (LOICV) to simulate the human-computer conversation. The difference between them is that, in LOOCV, the test case (target case) is taken away from the case base during the case retrieval process, while in LOICV, the test case is kept in the case base, and acted as the base case for the simulated retrieval process[4]. The LOICV has been successfully used in the CCBR community [20, 22].

The reason why we switch from LOOCV to LOICV lies in that:

- In the UCI case bases we use in LOOCV, many of the cases in a case base have the same solutions, so we can evaluate variant CCBR applications in a classification context. In this context, we can choose a case, which shares the same solution as the target case, as the base case of the target case. That is, it is possible to execute a simulated CCBR retrieval with the target case out of the case base.
- In the image processing software component case base, each software component has a unique solution (i.e., the software component itself). McSherry [21] refers to a case base with this property as an *irreducible* case base. The component retrieval problem is basically an identification problem rather than a classification problem. It is impossible to carry out a simulated CCBR retrieval with the target case being removed from an irreducible case base, as its unique solution is no longer represented in the case base.

In our KI-CCBR method, if we disable the four knowledge-intensive modules, Feature Inferencing, Knowledge-Intensive Question Ranking, Consistent Question Clustering, and Coherent Question Sequencing, the system becomes a knowledge-poor CCBR system (use only the statistical metric (occurrence frequency) to rank questions). Instead of enabling all these four modules at the same time, we enable them in a sequence of Feature Inferencing, Knowledge-Intensive Question Ranking, and Coherent Question Sequencing[5], respectively. With the above module enabling sequence, the average conversation session numbers needed to find the base case are 3.70, 3.64, 3.56, and 3.12, respectively, the latter with all modules enabled. That is, our knowledge-intensive CCBR method improves the efficiency by using 16% fewer conversation sessions (questions) to find the base case compared with the knowledge-poor CCBR method. Fig. 2

[4] The query-biased similarity method ensures that the test case is always included in the case group with highest similarity value, so the successful termination condition in LOICV, unlike that in LOOCV, is that the case group with the highest similarity value only contains the test case itself.

[5] In the simulated question-answering process, only the question with the highest priority is selected to be answered, so the enabling status of the consistent question clustering module has no influence on the evaluation results.

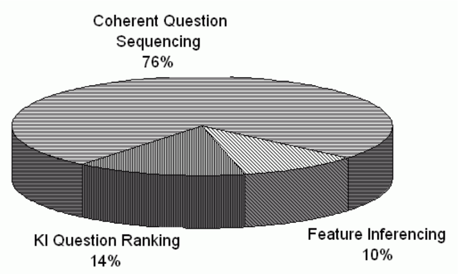


Fig. 2. Relative contribution of each KI-CCBR module to overall improvement in conversation efficiency

shows us that the relative improvements from Feature Inferencing, Knowledge-Intensive Question Ranking, and Coherent Question Sequencing are 10%, 14%, and 76%, respectively. The underlying reason why the coherent question sequencing module has such a major impact is that it guides users to answer the discriminative questions using a sequence ranging from general to specific and insisting on one description aspect instead of allowing a jump from one aspect to another which may be unrelated. However, the degree to which each module contributes to overall performance may depend on the different application domains and the contents of the knowledge bases.

6 Conclusion

In this paper, we note the difficulty of evaluating CBR systems using multiple case bases, and propose an evaluation strategy to use different data sources to assess different evaluation goals of a CBR system. First, all the evaluation goals are divided into two categories: domain-independent goals and domain-dependent goals. For domain-independent goals, we can choose many simple case bases and a statistical evaluation method for testing. For domain-dependent goals, we can choose one or a few target domain case bases and use multiple weak evaluation methods for testing. This evaluation strategy is applied to a knowledge-intensive conversational CBR method as a case study. The results of our case study indicate that such a combination of evaluation methods and test data sources can provide more solid evaluation results than is possible with a single evaluation method.

References

1. Simon, H.A.: Artificial Intelligence: an Empirical Science. Artif. Intell. **77** (1995) 95–127
2. Cohen, P.R., Howe, A.E.: How Evaluation Guides AI Research. AI Mag. **9** (1988) 35–43

3. Cohen, P., Howe, A.: Toward AI Research Methodology: Three Case Studies in Evaluation. Systems, Man and Cybernetics, IEEE Transactions on **19** (1989) 634–646
4. Cohen, P.R.: Empirical Methods for Artificial Intelligence. MIT Press, Cambridge, MA, USA (1995)
5. McSherry, D.: Interactive Case-Based Reasoning in Sequential Diagnosis. Applied Intelligence **14** (2001) 65–76
6. Aamodt, A., Plaza, E.: Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications **7** (1994) 39–59
7. Kolodner, J.: Case-Based Reasoning. Morgan Kaufmann Publishers Inc. (1993)
8. Santamaria, J.C., Ram, A.: Systematic Evaluation of Design Decisions in CBR Systems. In: Proceedings of the AAAI Case-Based Reasoning Workshop, Seattle, Washington (1994) 23 – 29
9. Díaz-Agudo, B., González-Calero, P.A.: An Architecture for Knowledge Intensive CBR Systems. In: Proceedings of the 5th European Workshop on Case-Based Reasoning. Trento, Italy (2000) 37–48
10. Aha, D.W.: Generalizing from Case Studies: A Case Study. In Sleeman, D.H., Edwards, P., eds.: Proceedings of the Ninth International Workshop on Machine Learning, Aberdeen, Scotland, UK, Morgan Kaufmann (1992) 1–10
11. Aamodt, A.: Knowledge-Intensive Case-Based Reasoning in Creek. In Funk, P., González-Calero, P.A., eds.: 7th European Conference on Case-Based Reasoning. Madrid, Spain, (2004) 1–15
12. Sørmo, F., Cassens, J., Aamodt, A.: Explanation in Case-Based Reasoning-Perspectives and Goals. Artificial Intelligence Review **24** (2005) 109 – 143
13. Newman, D., Hettich, S., Blake, C., Merz, C.: UCI Repository of Machine Learning Databases [http://www.ics.uci.edu/ mlearn/mlrepository.html] (1998)
14. Doyle, M., Cunningham, P.: A Dynamic Approach to Reducing Dialog in On-line Decision Guides. In: European Workshop on Advances in Case-Based Reasoning, Trento, Italy (2000) 49–60
15. Tong, X., Öztürk, P., Gu, M.: Dynamic Feature Weighting in Nearest Neighbor Classifiers. In: Proceedings of the 3rd International Conference on Machine Learning and Cybe (ICMLC2004). Volume 4., Shanghai, China, (2004) 2406 – 2411
16. Yang, Q., Wu, J.: Enhancing the Effectiveness of Interactive Case-Based Reasoning with Clustering and Decision Forests. Applied Intelligence, **12** (2001) 49 – 64
17. Bogaerts, S., Leake, D.: Facilitating CBR for Incompletely-Described Cases: Distance Metrics for Partial Problem Descriptions. In: Proceedings of the 7th European Conference on Case-Based Reasoning, Springer - Verlag (2004) 62–76
18. Bareiss, R.: The Experimental Evaluation of a Case-Based Learning Apprentice. In: the proceedings of the Case-Based Reasoning Workshop, Pensacola Beach, Florida (1989) 162 – 167
19. McLaren, B.M.: Extensionally Defining Principles and Cases in Ethics: an AI Model. Artificial Intelligence Journal **150** (2003) 145 – 181
20. Aha, D.W., Breslow, L., Muñoz-Avila, H.: Conversational Case-Based Reasoning. Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies **14** (2001) 9
21. McSherry, D.: Minimizing Dialog Length in Interactive Case-Based Reasoning. In: International Joint Conferences on Artificial Intelligence. (2001) 993–998
22. Gupta, K.M., Aha, D.W., Sandhu, N.: Exploiting Taxonomic and Causal Relations in Conversational Case Retrieval. In: European Conference on Case Based Reasoning, Aberdeen, Scotland, UK (2002) 133–147

23. Gu, M., Aamodt, A.: A Knowledge-Intensive Method for Conversational CBR. In Muñoz-Avila, H., Ricci, F., eds.: Case-Based Reasoning Research and Development, Proceedings of the 6th International Conference on Case-Based Reasoning. Chicago, Illinois, Springer Verlag (2005) 296–311
24. Gu, M., Tong, X., Aamodt, A.: Comparing Similarity Calculation Methods in Conversational CBR. In Zhang, D., Khoshgoftaar, T.M., Shyu, M.L., eds.: Proceedings of the 2005 IEEE International Conference on Information Reuse and Integration, Hilton, Las Vegas, Nevada, USA, (2005) 427 – 432
25. Gu, M., Aamodt, A.: Dialog Learning in Conversational CBR. To appear in the Proceedings of the 19th International FLAIRS Conference, Melbourne Beach, Florida, AAAI Press (2006)
26. Gu, M., Bø, K.: Component Retrieval Using Knowledge-Intensive Conversational CBR. To appear in the Proceedings of the 19th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Annecy, France, Springer (2006)
27. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. 2 edn. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers (2005)
28. Aha, D.W., McSherry, D., Yang, Q.: Advances in Conversational Case-Based Reasoning. Knowledge Engineering Review **20** (2006) 7
29. Göker, M.H., Thompson, C.A.: Personalized Conversational Case-Based Recommendation. In: the 5th European Workshop on Case-Based Reasoning(EWCBR 2000), Trento, Italy (2000)
30. Shimazu, H.: Expertclerk: A Conversational Case-Based Reasoning Tool for Developing Salesclerk Agents in E-Commerce Webshops. Artificial Intelligence Review **18** (2002) 223 – 244
31. Ferguson, A., Bridge, D.G.: Partial Orders and Indifference Relations: Being Purposefully Vague in Case-Based Retrieval. In: EWCBR '00: Proceedings of the 5th European Workshop on Advances in Case-Based Reasoning, London, UK, Springer-Verlag (2000) 74–85
32. Gu, M., Aamodt, A., Tong, X.: Component Retrieval Using Conversational Case-Based Reasoning. In Shi, Z., He, Q., eds.: Intelligent Information Processing II, Volume 163 of IFIP International Federation for Information Processing., Springer Science + Business Media Inc (2004)
33. McSherry, D.: Explanation in Recommender Systems. Artificial Intelligence Review **24** (2005) 179 – 197
34. Reilly, J., McCarthy, K., McGinty, L., Smyth, B.: Explaining compound critiques. Artificial Intelligence Review **24** (2005) 199 – 220