MID-CBR: An Integrative Framework for Developing Case-based Reasoning Systems
TIN2006-15140-C03

Enric Plaza *
IIIA - Consejo Superior de Investigaciones Científicas
Belén Díaz Agudo †
GAIA - Universidad Complutense de Madrid
Elisabet Golobardes ‡
GRSI - Universitat Ramon Llull

Abstract

Case-based reasoning (CBR) combines in an effective manner both learning from examples and usage of domain knowledge. Techniques for case retrieval and reuse should not be studied in an isolated manner; instead they should be designed and evaluated in a framework that integrates different types of CBR systems as proposed in this project.

Keywords: Case-based reasoning, lazy learning, soft computing, knowledge-intensive CBR systems, ontologies.

1 Project Goals

The main objectives of the project can be summarized as follows: (i) new ways to use techniques of soft computing for CBR, (ii) techniques for case reuse of a declarative and generic nature, (iii) techniques for case retrieval in knowledge-intensive CBR systems, (iv) integrating ontologies both in CBR systems and retrieval and reuse techniques (v) maintenance techniques both for case bases and for CBR systems capable of dealing with issues arising from design, implementation, and deployment of industrial strength CBR systems, including the empirical evaluation of the developed techniques by means of CBR prototypes implemented for several experimental domains, and (vi) developing component-based software platforms to support CBR systems development.

These goals are pursued by the tasks and subtasks of Figure 1, that also shows their time schedule and the output (Report or Software). These tasks organize the research activity

*Email: enric@iiia.csic.es
†Email: belend@sip.ucm.es
‡Email: elisabet@salleURL.edu
Concerning Subtask 1.1 (Data Intensive Retrieval), there has been a strong collaboration between the teams IIIA-CSIC and GRSI-URL, specially on the melanomas domain in cooperation with our EPO (external observer company), the Barcelona Clinic Hospital. We have studied the combination of Self-Organizing Maps (SOM) for clustering cases and organizing cases bases together with induction of their symbolic descriptions. This approach has been applied both to the melanomas domain and the intrusion detection domain, developed in collaboration with ISECOM, another EPO. Moreover, a methodology for helping expert how to efficiently perform the case retrieval from clustered case memory has been defined according to the data complexity and the performance desired. Finally, many works related to the automatic definition of similarity functions and cooperation schemes have been studied in order to improve the reliability of results in complex domains.

Concerning research in Knowledge Intensive Retrieval (Subtask 1.2) we have proposed a set of methods based on the recent DLs technological improvements. These ontology based retrieval methods have been used in different application domains. Namely, component configuration for templates retrieval, teaching for retrieval of behaviours, retrieval in case based planning and retrieval of recipes in the ICCBR-08 Computing Cooking Contest. Moreover, using argumentation in a setting with multiple CBR agents allows to elucidate the most relevant around types of techniques. These techniques are empirically tried out and evaluated in specific domains of application.

2 Achievement Degree of Project Goals

Concerning Subtask 1.1 (Data Intensive Retrieval), there has been a strong collaboration between the teams IIIA-CSIC and GRSI-URL, specially on the melanomas domain in cooperation with our EPO (external observer company), the Barcelona Clinic Hospital. We have studied the combination of Self-Organizing Maps (SOM) for clustering cases and organizing cases bases together with induction of their symbolic descriptions. This approach has been applied both to the melanomas domain and the intrusion detection domain, developed in collaboration with ISECOM, another EPO. Moreover, a methodology for helping expert how to efficiently perform the case retrieval from clustered case memory has been defined according to the data complexity and the performance desired. Finally, many works related to the automatic definition of similarity functions and cooperation schemes have been studied in order to improve the reliability of results in complex domains.

Concerning research in Knowledge Intensive Retrieval (Subtask 1.2) we have proposed a set of methods based on the recent DLs technological improvements. These ontology based retrieval methods have been used in different application domains. Namely, component configuration for templates retrieval, teaching for retrieval of behaviours, retrieval in case based planning and retrieval of recipes in the ICCBR-08 Computing Cooking Contest. Moreover, using argumentation in a setting with multiple CBR agents allows to elucidate the most relevant around types of techniques. These techniques are empirically tried out and evaluated in specific domains of application.
cases to be retrieved in a way that is both explicit and declarative. Concerning Textual and Conversational CBR (Subtask 1.4.) the novelty of our research is an approach to user-guided adaptation of texts based on combining IR techniques and clustering techniques. Concerning conversational CBR we have defined the model COBBER, that is based on ontologies and affective models to react to the user mood while interacting with the system.

Concerning research Retrieval and Reuse in Dynamic Environments (Subtasks 1.3 and 2.3), it has been focused on improving robot decision making in the RoboCup setting. Using CBR improved the previous strategy developed by Carnegie-Mellon University with which we’ve been collaborating. It’s worth noting the contribution to Retrieve techniques (by exploiting world symmetries) and to Reuse techniques (by adapting the outcome to the perceived situation and taking into account adaptation cost), as well as intensive experimentation with real robots.

Concerning Subtask 3.1 (Developing component-based software platforms to support CBR systems development) we have developed a new 2-layered architecture for jCOLIBRI 2. The top layer partially automates the generation of CBR systems through the composition of the elements in the framework. This composition is guided by templates that abstract the behavior of CBR systems and can be instantiated with the different components of the bottom layer. Concerning the Subtask 3.2, the design and maintenance of Case-Based Reasoning systems, we have investigated the use of introspective reasoning to improve the performance of Case-Based Reasoning (CBR) systems, in both reactive and proactive fashion, by guiding learning to improve how a CBR system applies its cases and by identifying possible future system deficiencies. We have also performed a case study on a library of templates to prototype Recommender Systems and we plan to extend it to other types of CBR systems Moreover, following the idea of working with cluster case-base memories, we have been working with the maintenance of this kind of organizations. Concerning Subtask 3.3, Evaluation techniques for CBR systems, we have studied and analyzed the related works in this area in order to propose an experimental methodology for allowing non experts in this topic the statical analysis of results of their machine learning systems.

Concerning Task 4.1 (Coordination) the organization of yearly plenary workshops in addition to task-specific meetings convened by the responsible researcher has worked as established in the project proposal. Regarding Prototypes evaluation (Subtask 4.1) jCOLIBRI 2 is available from August 2007 and has been downloaded 5000 times from more than 70 different countries. Our next goal now is to get contributions from other CBR researchers. There are also some commercial applications using jCOLIBRI, namely KoBas: knowledge management system built by the consultants Wittmann & Partner in Rumania; and EMEF (Portuguese Railway Company) gave its 1st R&D Award to an intelligent system for failure diagnostics developed in collaboration with the University of Porto. Concerning relationship with external observer companies we have had a fluid collaboration and a satisfactory feedback from them, resulting in research actives that have lead to publications in the domains of medicine (melanoma analysis with the Barcelona Clinic Hospital), music personalization (with MusicStrands.com) and computer network intrusion (with ISECOM).

3 Assessment of Project Outcome

Concerning Task 4.1 (Coordination), we remark that the annual plenary workshop in addition to task-specific meetings has worked satisfactorily in creating synergies and team cohesiveness.
3.1 Assessment of IIIA-CSIC Project Outcome

Concerning Subtask 1.1 (Data Intensive Retrieval), one of our focus has been generating explanation to the user of the CBR system: in [3] we analyzed several usages of generalization commonly done in both lazy and eager learning methods. An application of generalizations as explanation of the retrieval was that introduced in [2] were the goal was to explain the classification of chemical compounds as carcinogenic and no carcinogenic. This application was specially useful since because the classification of carcinogenic compounds is not clear, experts can use the explanations to decide the final classification of a compound. An extension of this work was also published in [7]. In [4] we proposed to store the explanations and use them as partial concept descriptions. A second focus, in collaboration with GRSL-URL, we analyzed the combination of SOM and cluster explanations on standard domains (such as those of the UCI Repository) [23]. This approach was applied to the domain of detecting intrusions in a computer network [14], and showed that the experts could better understand the system with the symbolic explanation of clusters [15]. These approach was applied to the classification of melanoma in situ of the Barcelona Clinic Hospital, using CBR to help discover new ways to cluster data [26] and proposing new ways to classify them [5, 6]. Finally, we have proposed a new research line on how to reuse human experience available through the web in [53, 56].

Concerning Task 1.2 (Knowledge Intensive Retrieval), we have focused on the improvement of the retrieval process using an argumentation framework in a multi-agent setting. Our work show that agents that can justify (explain symbolically) their prediction can engage in argumentation over the correctness of possible solutions, the reasons pro and against, and the counter-examples that support such discussions; as a result we showed that (1) the individual CBR agents can learn from the argumentation with very few examples and (2) both the individual and group accuracy improves when using argumentation [12, 50, 51, 52, 53, 54, 57].

Concerning research Retrieval and Reuse in Dynamic Environments (Subtasks 1.3 and 2.3), international relationship has been embodied in a strong cooperation with the RoboCup team of Manuela Veloso at Carnegie Mellon University, while the contributions have been rise to 1 finished PhD (Raquel Ros, March 1st 2008, “Action Selection in Cooperative Robot Soccer using Case-Based Reasoning”) and 1 PhD in the writing (Arnau Ramisa), both supervised by R. López de Mántaras. Several scientific publications report this research in competitive international conferences [78, 77, 74, 71, 60, 61, 62] (including the best conference paper award for [78]) and in two journals [76, 75].

Concerning Task 2.1, Reuse for configuration tasks, we have studied, in collaboration with GAIA-UCM, the use of declarative reuse techniques like constructive adaptation for originality-driven tasks (ODT) [20] — i.e. for tasks were a solution has to be a novelty when compared to preexisting ones. We showed that constructive adaptation improves novelty over classical transformational adaptation techniques. Concerning Task 2.2 (Reuse for planning and scheduling) we have focused on developing techniques that customize a sequential solution to a group of users. Specifically we worked in collaboration with our EPO MusicStrands on the domain of musical programming, developing Poolcasting, an automatic DJ that schedules songs customized to the tastes of the members of an audience [8, 9, 10, 11, 54]. The Poolcasting generalizes the concept of “case reuse” to include the target of the solution (i.e. the audience) and for this reason received the Best CBR Application Paper Award at the 2008 Int. Conf. on Case-Based Reasoning [8]. Claudio Baccigalupo’s Ph.D. is expected to be defended In June 2009 at the Univesitat Autònoma de Barcelona.
Concerning the Task 3, the design and maintenance of Case-Based Reasoning systems, the goals of the task have been achieved according to the working plan: The proactive method is the core of the CB-Inspector prototype [43, 44, 45], and has been applied to the RoboSoccer domain (Subtasks 1.3 and 2.3) and showed how founding dubiosity regions helped us to analyze the case-base of that CBR system. The reactive model has been successfully incorporated in to the Auto-CBR prototype. We have shown our model’s benefits with experimental results from tests in an industrial design application ([1]). Both methods have been published at international research forums. The introspective model, developed with collaboration of Prof. David Leake (Indiana University, USA), will be described in a chapter of the book “Metareasoning: Thinking about Thinking” (The MIT Press). The PhD student Oguz Mulayim has achieved the DEA degree in 2008 [42].

3.2 Assessment of GAIA-UCM Project Outcome

The main results from the UCM group have been built around jCOLIBRI 2, an object-oriented framework in Java for building CBR systems based on the reuse of previously developed CBR systems components. The first version of jCOLIBRI [70] is a complete CBR development architecture supporting many features like graphical interfaces, Description Logics and ontologies, Textual CBR, evaluation, among others. The complete work has been published in [19]. jCOLIBRI 2 defines a two-level architecture. The bottom layer is oriented to java developers and provides an architecture that is able to integrate diverse CBR reusable methods in a library. This layer provides the basic blocks for building CBR systems that can be easily extended and reused by programmers. jCOLIBRI 2 was released in August 2007 and since then it has served as the integration framework to develop CBR systems, has allowed the integration of different research lines, has been the main diffusion way for our research results and the way to collaborate with other national and international research groups that have incorporated new methods to the framework.

Concerning Subtask 1.2 (Knowledge Intensive Retrieval) see [68] and [18] for a detailed description of the main knowledge intensive retrieval techniques based on DLs technology. These new methods are available in jCOLIBRI2. We have also research on the use of knowledge intensive methods to help reasoning to build natural language interfaces to CBR systems [21]. These methods have been applied in different domains. Namely templates retrieval [63, 67], retrieval of videogame behaviours represented as state machines [58, 59], retrieval in case based planning [50], and retrieval of recipes in the computing cooking contest [17].

Concerning Subtask 1.4 Textual CBR (TCBR) the work of the UCM group in textual CBR has been built in collaboration with Nirmalie Wiratunga and her group from Robert Gordon University in Scotland. In [65] the CBR textual methods have been used in Challenger 1.0 a TCBR application that addresses the Air Investigation Reports challenge proposed in the ICCBR 2007 workshop program. Textual methods have been also used in the Computer Cooking Contest application [17], and in a tutor system to answer questions in the videogame JV2M. Contributions regarding conversational CBR are the main result of the Ph.D. thesis of Hector Gomez Gauchia [38] on COBBER, a model and an architecture to design conversational systems [39].

Research in Subtask 2.1 (reuse in case based configuration) builds upon previous work of the IIIA and the UCM groups. As the main contribution in this subtask we have studied the reuse task in CBR systems from the result originality point of view. The work presented
TIN2006-15140-C03

in [20] analyzes the role of reuse in CBR systems in originality driven tasks (ODT). These reuse methods have also been applied in two experimentation domains: reuse of semantic web services [69, 72, 36, 73] and templates based design domain [64, 67].

Concerning Subtask 2.2 (Reuse in planning and scheduling) we have studied how well tested and reusable knowledge and techniques from the Knowledge Representation arena can be integrated into a hierarchical planner. As the experimentation domain we have used strategy board games [80] and composition of Semantic Web Services [79]. The former has been built in collaboration with Hector Muñoz Avila and his group from University of Leigh in Maryland (USA), as the result of a three months visit from the UCM researcher (Antonio Sanchez Ruiz-Granados) to Leigh by February to April 2007. In [81] we have proposed a method of adaptation through Planning in Knowledge Intensive CBR and have applied this method in the teaching domain using games to retrieve and reuse exercises for a student taking into account the student characteristics and his/her previous knowledge.

jCOLIBRI 2 is an object-oriented framework in Java for building CBR systems based on the reuse of previously developed CBR systems components. A large effort of GAIA-UCM has been devoted to developing component-based software platforms to support CBR systems development (Subtask 3.1) based on jCOLIBRI 2 (presented in December 2008 as the PhD Thesis of Juan Antonio Recio García [63]). A CBR ontology is used to formalize syntactical, semantical and pragmatics aspects of the reusable components of the framework [67]. In [64, 66] we have presented our work on template based design of CBR systems using as case study the prototyping of Recommender Systems in jCOLIBRI. Recommender Systems Templates have been built in collaboration with professor Derek Bridge, from University College Cork in Ireland as the result of a two months visit of the UCM researcher Juan A. Recio-Garcia to Cork University (Ireland).

3.3 Assessment of GRSI-URL Project Outcome

Concerning Task 1.1 (Data Intensive Retrieval), we have focused on the context of classification problems where the cases are complex, with imprecise knowledge and with large case bases, that is, in problems with the traditional methods may not offer the expected results. For this reason, we have tackled this issue from the point of view of soft-computing because this kind of techniques allow a better management of knowledge than traditional hard computing techniques. On the one hand, we have used the searching capabilities of Genetic Programming and Grammar Evolution with the aim of automatically defining similarity functions ad hoc a domain. This is specially useful for domains where the conventional distances do not offer the expected results due to the domain geometry and/or its structure [28, 27]. The same idea has been applied for defining case retrieval strategies from ensemble systems [30] and in a more knowledge intensive approach using the melanoma domain of the HCPB [48, 47]. On the other hand, we have worked with Self-Organizing Maps in order to organize the case memory in clusters and, consequently, try to take profit from the patterns identified in order to use only the potentially useful cases. Although this should reduce the computational time without reducing the accuracy, the data complexity may difficult the pattern identification and system could not find the right solutions. This problematic has motivated the definition of a methodology for helping expert to decide how many clusters and cases have to be used according to the data complexity in order to achieve the desired performance in terms of time and accuracy [31, 32, 41, 40, 22]. The mechanisms used to assess the data geometry are
the complexity measures. This work promoted the relation with IIIA-CSIC as the section 3.1
described for generating symbolic explanations in order to help expert to understand better
the case retrieval results in the melanoma domain [23, 25, 26, 22] and in the telematic domain
[15, 34, 14, 16]. Moreover, we have also worked in defining new case retrieval strategies based
on symbolic descriptions [24] and in how to improve the reliability of the reuse phase through
the definition of confidence levels from the patterns [33].

Concerning Task 3.2 (Maintenance), the work has been developed at the same time than
Task 1.1 due to the closer relation with the cluster case-based organization. In this sense, we
have worked in the maintenance of the case memory in terms of allowing the incremental and
semisupervised learning in SOM [29] even this technique does not allow it. Most of the prece-
dent tasks in the project imply the development of new methodologies in the different phases of
a CBR system. Although the project is focused on the development of these methodologies, it
seems obvious that a test of the results obtained on well-known cases and new proposed prob-
lems is needed. For this reason, we have defined an experimental methodology for analyzing
the results of machine learning systems based on the different types of experiments and statist-
ics test [11]. All this work is placed in the context of the Task 3.3 (Evaluation). Concerning
Task 4.2 (Evaluation prototypes), we have started to work with UCM for integrating all the
cluster case-based algorithms inside the framework jCOLIBRI developed by them. Concerning
Task 4.3 (EPO Relationship), the cooperation with both EPOs, ISECOM and HCPB, has
been a successful because many works have been done in cooperation. Moreover, the results
are positive because many works have been accepted in conferences and journals related to the
telematic domain [15, 14, 16, 34] and the melanoma domain [22, 18, 47, 25, 26] along 2008.
Moreover, the relation with the HCPB and ISECOM has deep contributed to the delivering of
two DEAs [16, 13] and one Ph.D. will be presented in the next months.

References
Based Reasoning Research and Development, ICCBR-07, number 4626 in Lecture Notes in Arti-
24, 2008.
R. Weber, K. Bach, and M. Petridis, editors, Procs. of Workshop on Uncertainty, Similarity and
Gruart, and C. Torras, editors, Artificial Intelligence: Research and Development (CCIA-2007),
[7] E. Armengol. Using explanations for determining carcinogenicity in chemical compounds. Interna-


