Evaluating an Intelligent Diagnosis System of Historical Text Comprehension

Grammatiki Tsaganou[a], Maria Grigoriadou[a], Theodora Cavoura[b], Dimitra Koutra[a]

[a]University of Athens, Department of Informatics and Telecommunications, GR-15784, Athens, Greece,
e-mails: gram@di.uoa.gr, gregor@di.uoa.gr, mik@altec.gr
[b]University of Thessaly, Dept. of Education, Argonafton & Filellinon strs, GR-38221, Volos, Greece, e-mail: theokav@pre.uth.gr

Abstract

This work aims to present and evaluate a Fuzzy-Case Based Reasoning Diagnosis system of Historical Text Comprehension (F-CBR-DHTC). The synergism of fuzzy logic and case based reasoning techniques handles the uncertainty in the acquisition of human expert's knowledge regarding learner's observable behavior and integrates the right balance between expert's knowledge described in the form of fuzzy sets and previous experiences documented in the form of cases. The formative evaluation focused on the comparison of the system's performance to the performance of human experts concerning the diagnosis accuracy. The system was also evaluated for its behavior when using two different historical texts. Empirical evaluation conducted with human experts and real students indicated the need for revision of the diagnosis model. The evaluation results are encouraging for the system's educational impact on learners and for future work concerning an intelligent educational system for individualized learning.

Keywords: Fuzzy Case based reasoning; Historical text comprehension; Diagnosis evaluation; Learner model.

1 Introduction

In an Intelligent Tutoring System (ITS), learner diagnosis process imitates the human expert's process of inferring the student's internal characteristics from his observable behaviour (VanLehn, 1988). In the domain of comprehension of history, this computational diagnostic process imitates a human expert's ability to estimate how the learners comprehend the historical text. An attempt towards this direction is our previous work concerning the Learner Model of learner's cognitive profiles for Historical Text Comprehension (LMHTC) (Tsaganou et al, 2003). Such a learner model demands knowledge acquisition from a human expert, which means knowledge extraction regarding the student's observable behaviour, its complete and accurate description and transfer to a knowledge base and sometimes to an inference engine. The main obstacle in this process is the uncertainty derived not only from the knowledge communication among the developer, the human expert and the system, but also from inaccuracy of the information captured and of approximation involved in all process steps (Richter, 2001).

Case Based Reasoning (CBR) is claimed to be a paradigm that is more akin to the human way of solving complex diagnostic problems in domains like medicine or law. A human expert solves a diagnostic problem using rules derived from his previous experience-cases, whereas a novice requires complete and concrete rules. CBR integrates the right balance between hard to acquire expert knowledge and more easily acquired knowledge in the form of cases. So, for building of an ITS, CBR helps more easily than other methods to overcome problems of knowledge acquisition from the expert. CBR has been proposed for a variety of diagnostic applications (Kolodner, 1993), has been used in educational systems such as in CELIA, for modelling the memory and reasoning capabilities of a novice (Kolodner, 1993), in Engines for Education for case based coach (Schank et al, 1994), in Tutoring and Help Systems (Weber et al, 1998; Burke et al, 1994), in SYIM for distance education (Tsinakos et al, 2001), as part of the student modelling process in ITS systems (Shiri et al, 1998).

For overcoming complex problems, like uncertainty in knowledge acquisition, developers recently build more hybrid case-based and knowledge-based systems than pure CBR systems. Fuzzy logic is designed to operate with linguistic expressions and express imprecision and subjectivity in human thinking. Fuzzy logic contributes to CBR in overcoming problems of managing the uncertainty and problems concerning case adaptation by improving performance of case retrieval (Jeng et al, 1995). In research community the interests of fuzzy logic and CBR for diagnosis recently intersect (Dubois et al, 1997; Richter et al, 1999; Hansen, 2000). Fuzzy logic is widely used in student modelling where variables are continuous, imprecise, or ambiguous. Fuzzy-based techniques have been used in educational systems for flexible case-based querying (Calmes et al, 2002).

Often, we think of evaluation in terms of how well AI systems, perform. Evaluation of a student model tries to answer a question that is central to cognitive science, AI and education (Littman et al, 1988): What is the relationship between the architecture of the student model and its behavior? The evaluation methods can be used to construct an accurate picture of this relationship. The issue of noise, for example, must be addressed by any realistic system performing pedagogical diagnosis (Wenger, 1994). Three sources of noise are considered: First, noise in the data, which means non-consistency of student's behavior over the time. Second, noise in the diagnostic process, which means inherent ambiguities in the diagnostic process. Third, noise in the model of communicative knowledge, which means that

noise itself can become a source of information. Evaluating a CBR system involves two separate processes, called verification and validation (Watson, 1997). Verification is concerned with building the system right, ensuring that the system gives correct answers. Validation is about building the right system, the one that the users want. The notion of evaluation, in this work, is seen as evaluating how close the diagnostic results of a Fuzzy-Case Based Reasoning system are to that of experts (Althoff, 1994; Althoff, 1995; Chin, 2001).

In this contribution we use Fuzzy Case-Based Reasoning technique for the construction of a system for Diagnosis of students' cognitive profiles of Historical Text Comprehension. In Section 2 the underlying learner model LMHTC is presented. In Section 3, the F-CBR-DHTC system is described. In Section 4, the user interface, system implementation and preparation for use are presented. In section 5, we focus on the evaluation by human experts of system's diagnosis accuracy and of its performance when using a new historical text. We also describe the evaluation results and how they were used for the revision of the diagnostic model. In section 6, we conclude, present lessons learned and give links to ongoing research.

2 The Learner Model of Students' HTC

2.1 Models of HTC

Comprehension of historical text is a special kind of the complex and interactive cognitive process (Briton , 1996). Historical text is defined as a causal and transformational system. It is characterised as transformational because it describes the representation of a historical change, which has happened in a particular place and time. It is characterised as causal because it describes an historical occurrence, which is interpreted by a series of causal links.

The reader utilises certain fundamental cognitive categories for establishing and organising the meaning of the text (Baudet et al, 1992). During comprehension the reader attributes meanings to causal connections between occurrences in the historical text (Cavoura, 1994). In the level of comprehension as a cognitive task, the learner composes a representation of the historical text. This representation is a system, which contains the cognitive categories: *event*, *state* and *action*. Comprehension of the historical text is associated with causal connections and arguments made by the reader. The arguments are based on the three cognitive categories. For the interpretation of the learner's cognitive processes we analyse his discourse tracing the recognition or not of the three cognitive categories.

2.2 The LMHTC

The underlying model of the F-CBR-DHTC system is the LMHTC model, which is based on MOCOHN, a pencil-and-paper diagnosis model, and on further experimental research (Cavoura, 1994; Tsaganou et al, 2003). The LMHTC presents to the learner a historical text in the appropriate form and question-pairs with alternative answers. The historical text includes factors/ instances, which represent the 3 cognitive categories action, state and event. For every factor a question-pair is submitted to the learner. Each question-pair consists of two questions concerning the same factor. The first question in the question-pair is relative to the learner's position about the significance of this factor and the second question is relative to the learner's the justification of this position. The learner has to use the given alternative answers, which correspond to position and justification, in order to express his position for certain historical issues and support it by selecting a justification. The answers concerning position and justification are classified as scientific, towards-scientific and non-scientific, as they are depicted in Table 1 and Table 2. The alternative answers reflect scientific thought, towards acquiring scientific thought, and non-scientific thought.

Table 1
Classification of answers concerning position

Position	Answers
Scientific	learners attribute minimum importance to events and states and maximum or medium importance to actions
towards-scientific	learners attribute medium importance to events and states
non-scientific	learners attribute maximum importance to events or minimum importance to actions

Table 2
Classification of answers concerning justification

Justification	Answers
Scientific	learners ground their answers up on the scientific historical thought
towards-scientific	learners base their answers on the *common sense* schemas expressing experience, quantity, continuity and attitudes, which means learners are towards acquiring scientific thought
non-scientific	learners give cyclic answers based on recycling the questions and consequently they refer to non- scientific thought

Figure 1 depicts a historical text concerning 5 different factors of the outbreak of French Revolution. In the historical text, one factor represents the cognitive category *event*, one the *state* and three factors represent the *action*.

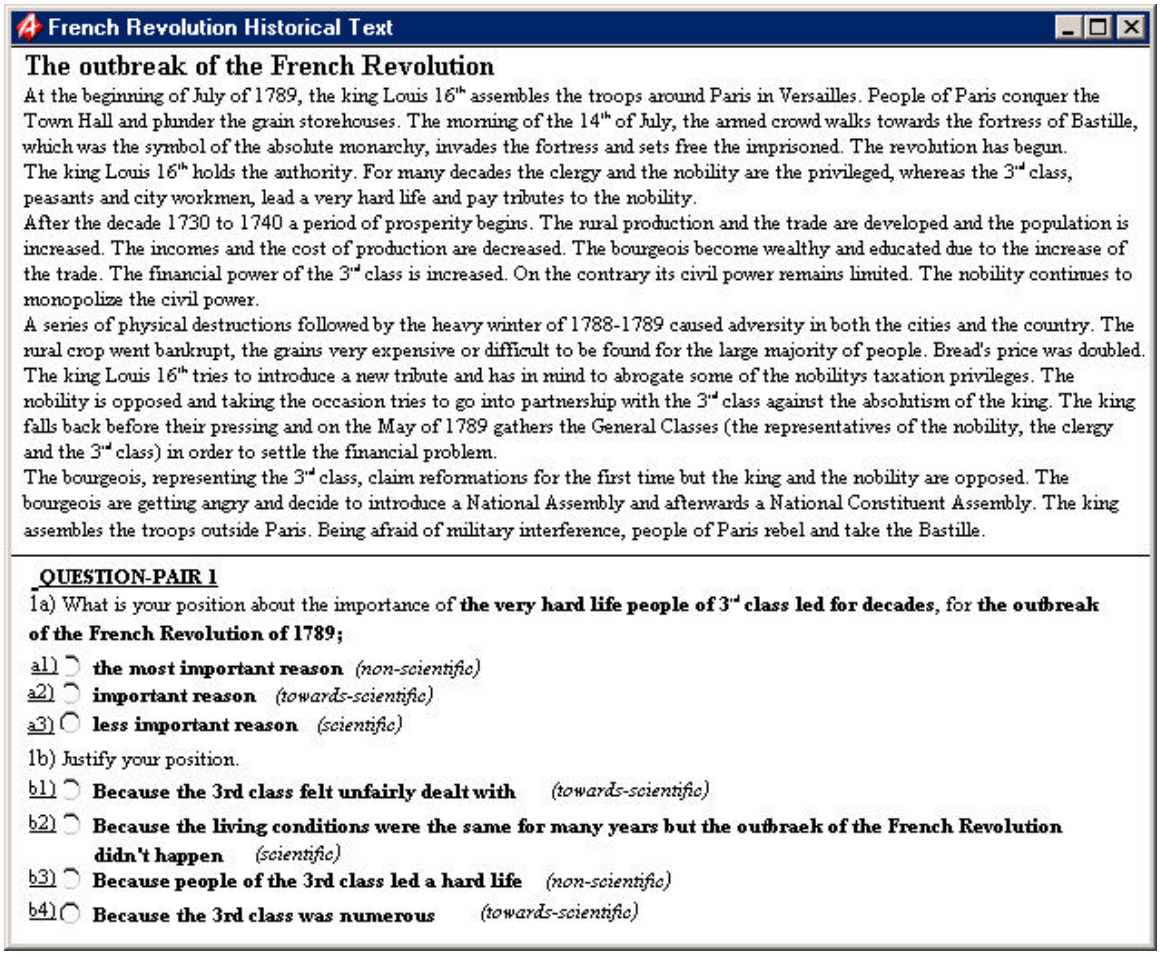


Figure 1: A screenshot depicting a historical text concerning the outbreak of French Revolution, question-pair number 1, alternative answers and characterizations of the answers.

Examples of alternative answers considered, which refer to positions and to justifications are also depicted in Figure 1. The answers a1 to a3 are alternative answers to question 1a concerning the position, whereas the answers b1 to b4 are alternative answers to question 1b concerning the corresponding justification. Figure 2, also indicates the characterisations of the answers, which are not visible to the learner. Answer a1 is *non-scientific*, a2 is *towards-scientific* and a3 is *scientific* answer. Answer b1 is *towards-scientific* expressing experience, b2 is *scientific*, b3 *non-scientific* and b4 is *towards-scientific* expressing quantity, see Table 2.

2.3 Arguments

For every question-pair the combination of the learner's position and the corresponding justification constitute the learner's *argument*. An *argument* is defined as *complete* when both *position* and *justification* are scientific. Otherwise the argument is *non-complete*. The expert defines the different degrees of *argument completeness*. The argument completeness, which is associated with the recognition or not of an instance of a cognitive category, is used as a vehicle to reveal the degree of the recognition or not of the corresponding cognitive category. Table 3 demonstrates all possible combinations of position-justification pairs, the corresponding argument completeness and characterization and the degree of recognition of a cognitive category. Possible values of the *argument completeness* are: *complete, almost complete, intermediate, nearly incomplete and incomplete*. For the characterization of argument completeness justification weights more than position does. So for example, towards-scientific position and scientific justification result in almost complete argument, whereas the opposite one that is scientific position and towards-scientific justification result in nearly incomplete argument.
Argument completeness describes the learning difficulties of the learner concerning the cognitive categories, which the learner does not recognize. This qualitative characteristic of the arguments reflects the degree of recognition of a cognitive category that is the degree of comprehension of the historical text.

Table 3
Argument completeness values concerning position -justification combinations.

Position	Justification	Argument Completeness	Argument characterization	Status of recognition of the cognitive category
scientific	scientific	complete	scientific	recognition
towards scientific	scientific	almost complete	towards-scientific	towards-recognition
non-scientific	scientific	intermediate	towards-scientific	towards-recognition
scientific	non-scientific	nearly incomplete	towards-scientific	towards-recognition
scientific	towards-scientific	nearly incomplete	towards-scientific	towards-recognition
towards-scientific	towards-scientific	nearly incomplete	towards-scientific	towards-recognition
non-scientific	towards-scientific	incomplete	non-scientific	non-recognition
towards-scientific	non-scientific	incomplete	non-scientific	non-recognition
non-scientific	non-scientific	incomplete	non-scientific	non-recognition

2.4 Classification of the cognitive categories

Historical actions constitute the core of the historical discourse. According to relevant research, during the comprehension of the historical text, the recognition of the cognitive category action is more important than the recognition of the cognitive category state (Cavoura, 1994). The recognition of the cognitive category event is less important than the recognition of the cognitive category state. This results in the classification of the cognitive categories. Table 4 demonstrates the quality values of the cognitive categories. Possible values of the quality are *superior* for the action, *medium* for the state and *inferior* for the event.

Table 4
Quality values of the cognitive categories

Cognitive category	Category quality
action	superior
state	medium
event	inferior

2.5 The Cognitive Profiles of HTC

The status of recognition of the three cognitive categories: event, state and action are used to formulate the cognitive models of HTC, which reflect the learners' levels of historical thought (Tsaganou, 2002). The learner's cognitive profile of HTC is formulated taking into account the number of his complete arguments. The cognitive profile expresses the recognition or not of the cognitive categories. Table 4 depicts the cognitive models and the cognitive profiles of HTC. The general categories of cognitive models considered are *Historical Thought* (HT), *Towards Acquiring Historical Thought* (TAHTnx) and *Non-Historical Thought* (NHT). TAHTnx cognitive models are categorised in more detail according to the number n of recognised by the learner cognitive categories and to the number x of their instances in the historical text. TAHT1 means that the learner recognises 1 instance of a cognitive category, whereas TAHT1x means that the learner recognises x instances of a cognitive category, where x>1. The same stands for TAHT2, TAHT2x, TAHT3 and TAHT3x. The numbers n and x are used for the formulation of the learner's cognitive profile.

Learners with *Very Low* profile seem to have serious difficulties in thinking historically. Learners characterized by terms like *low, nearly low, below intermediate, above intermediate, nearly high and high,* seem to encounter difficulties in thinking historically. Learners with *very high* profile seem to have no learning difficulties in thinking historically.

2.6 The Profile Descriptor

The *profile descriptor* describes the learner's cognitive profile and denotes different perspectives of the profile. The learner's *profile descriptor* is formulated taking into account all of his arguments which may have different degree of completeness. The *profile descriptor* carries detailed description pertaining to the *quality* of the cognitive categories and the completeness of the *arguments,* which are attached to every cognitive profile. The *profile descriptor*, which models uncertainty associated with observation, depicts the learner's problems in the recognition of the cognitive categories and reflects his learning difficulties.

For example, selection of the answers a3 and b2 of figure 1 constitutes a *complete argument* of *medium category* and indicates the recognition of one instance of the cognitive category (in this example the category *state*). Selection of the answers a2 and b4 constitutes a *nearly incomplete argument* of *medium category,* which indicates non-recognition of the cognitive category.

A *nearly low* cognitive profile of a learner, during comprehension of a historical text with 5 factors and 5 corresponding question-pairs, can be accompanied by the following *profile descriptor*: "The learner gives *one complete argument of inferior category, one nearly incomplete argument of superior category, one nearly incomplete argument of superior category, one incomplete argument of superior category and one incomplete argument of medium category".*

Table 5
Cognitive models, number and types of recognised cognitive categories and corresponding cognitive profiles, n in {1,2,3} and x>1.

Cognitive models	Number and types of recognised by the learner cognitive categories	Cognitive profiles
NHT	no event or state or action	very low
TAHT1	1 event or 1 state or 1 action	low
TAHT1x	more than one events or more than one states or more than one actions	nearly low
TAHT2	(1 event and 1 state) or (1 event and 1 action) or (1 state and 1 action)	below intermediate
TAHT2x	(more than one events and states) or (more than one states and actions) or (more than one events and actions)	above intermediate
TAHT3	1 event, 1 state and 1 action	nearly high
TAHT3x	more than one events, states and actions	high
HT	all events, states and actions	very high

3 The F-CBR-DHTC System

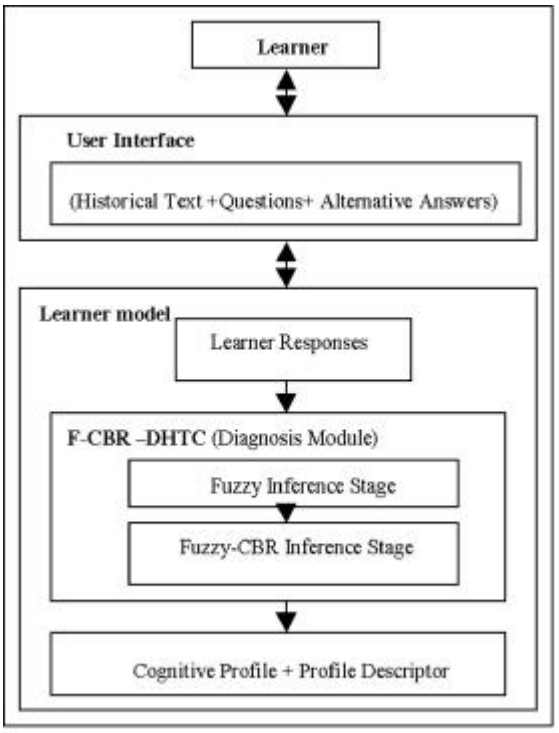


Figure 2. Structure of the F–CBR-DHTC system

The F-CBR-DHTC system is a hybrid Fuzzy-CBR system, which implements the diagnostic module of the LMHTC. It handles the complexity of diagnosis of student's cognitive profile and profile descriptor. The system encourages the learner to read the historical text and answer to question-pairs selecting from the alternative answers. The learner's responses define his observable behavior. The F-CBR-DHTC system, Figure 2, solves the diagnostic problem in two stages: (1) the *Fuzzy inference stage*, which infers the arguments' completeness reflecting the degree of recognition of the cognitive categories and (2) the *Fuzzy-CBR inference* stage, which infers the learner's cognitive profile and profile descriptor.

3.1 The Fuzzy Inference stage

The *Fuzzy inference stage* uses fuzzy rules, which incorporate the description of expert's knowledge concerning the student's answers to questions (position-justification pair values) and infers the argument completeness. This stage also expresses argument completeness with fuzzy sets, which are used in the next stage for defining similarity values.

In this stage the symbolic input data (student's responses) are transformed into linguistic terms (Jeng et al, 1995). $A=\{A_1, A_2,..., A_n\}$ and $B=\{B_1, B_2,..., B_n\}$ are the types of responses: where $A_1, A_2,..., A_n$ concern *position*, and $B_1, B_2,..., B_n$ concern justification. The term set of A_n is $T(A_n)=\{D_{n1}, D_{n2},..., D_{nk}\}$ with k linguistic values and the term set of B_n is $T(B_n)=\{E_{n1}, E_{n2},..., E_{nl}\}$ with l linguistic values. The sets $T_A= \{T(A_1), T(A_2),...,T(A_n)\}$ and $T_B=\{T(B_1), T(B_2),...,T(B_n)\}$ are the fuzzy sets of all term sets that represent the observable behaviour. The symbolic input is fuzzified by means of the linguistic variable values D_{nk} and E_{nl}. The numbers k, l, D_{nk} and E_{nl} are defined by the developer with the help of the human expert. Student's argument *completeness* C_n, result as combinations of the independent input data: *position* and *justification* and is represented as linguistic variables with m linguistic values and corresponding term set: $T(C_n)=\{C_{n1}, C_{n2}, \ldots, C_{nm}\}$. Let consider an example, of a historical text with n=5 arguments, k=3 linguistic values concerning position, l=3 linguistic values concerning justification and m=5 linguistic values concerning argument *completeness*. Then the sets are A_1= *position for event1*, B_1= *justification for event1* and C_1=*argument for event1* and the term sets are $T(A_1)$= T(*position for event1*)$=\{D_{11}, D_{12}, D_{13}\}$ ={right, mediocre, wrong}, $T(B_1)$= T(*justification for event1*)$=\{E_{11}, E_{12}, E_{13}\}$= {right, mediocre, wrong}, $T(C_1)$=T(*argument for event1*)$=\{C_{11},C_{12},C_{13},C_{14},C_{15}\}$= {incomplete, nearly incomplete, intermediate, almost complete, complete}.

IF A_n is D_{nk} AND B_n is E_{nl} THEN C_n is C_{nm} (1)

Using fuzzy rules of the form (1), which take into account position and justification values, this stage infers the corresponding argument completeness. Each fuzzy rule results in one fuzzy set. Figure 3, demonstrates values of the membership functions that express to what degree the values in the universe of discourse belong to the fuzzy set "how *complete* an *argument for event1* is".

The fuzzy sets are used by the next stage of the system to express local similarity measures. The output of the *fuzzy inference module* is the n-dimensional vector S_n, in [0,1], of student's argument characteristics S_n, which constitutes the input data to the *Fuzzy-CBR inference module*.

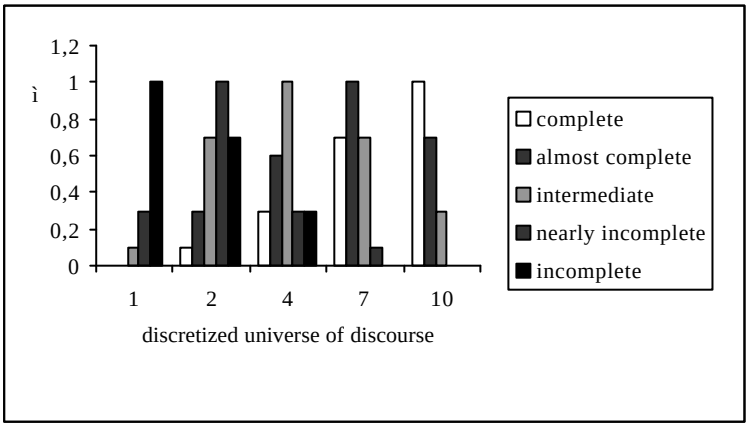


Figure 3. Discrete fuzzy sets for the inferior category *argument for event1* with respect to "how *complete* the *argument for event1* is"

3.2 The Fuzzy-CBR Inference stage

The aims of *Fuzzy-CBR inference stage* are to define case structure using the argument characteristics, define similarity measures based on the fuzzy sets, which have been defined in the previous stage, and proceed to the case retrieval and case adaptation process (Bergmann et al, 1998; Leake, 1996). In the *Fuzzy-CBR inference stage* the learner's behavior represented by the argument completeness constitutes the corresponding case. A case is viewed as a set of attributes, which contains (apart from the case name) two non-empty sets (Dubois et al, 1997): the problem description attributes subset S (argument completeness) and the solution description attributes subset T (cognitive profile and profile descriptor). Expressed in terms of the fuzzy-CBR framework, the case is declared in the following form: <casename, S_1, S_2,..., S_n, T_1, T_2>. In this stage, reasoning steps are based on the hypothesis that similar problems have similar solutions.

The challenging problem is to determine the degree to which stored cases are similar (Aamodt et al, 1994). The local similarity measures between argument characteristics of two cases are calculated according to the previously defined fuzzy sets. The global similarity between two cases is computed using a fuzzy knn algorithm. A fuzzy knn algorithm facilitates the retrieval process from the case base, handles case adaptation by exploiting the defined as fuzzy sets similarity values between the arguments characteristics (Hansen, 2000). A learner's cognitive profile and profile descriptor are inferred by the *Fuzzy-CBR inference module* after comparisons of the *arguments* of a case with the corresponding *arguments* of the other learners stored in a case base.

The solution description attribute subset $T= \{T_1, T_2,\ldots,T_n\}$ of the problem description vector S of current case q contains the attribute-values and T is the diagnosis - the requested value of *cognitive profile* and the *profile descriptor* (Aamodt,

1995; VanLehn, 1988). The retrieval process finds a case with an information vector S' most similar to S. Once a matching case is retrieved the system it is used to form the solution T.

4 System Preparation for Use

4.1 The environment of the F-CBR-DHTC system

The F-CBR-DHTC system has been implemented using an object-oriented language (C++) combined with a powerful graphical interface builder (Toolbook Instructor). The implementation method was chosen for its ability to handle a case base and its flexibility in solving problems. The environment of F-CBR-DHTC provides the users with an easy-to-use interface through which (a) the learners are given the historical text to read and questions with alternative answers to select according to their opinion and (b) the experts or the history teachers can author a novel environment by adding a new historical text with questions an alternative answers.

4.2 Case Base Initialisation

For the case base initialisation was used a sample of 60 most representative cases of the problem domain. We experimented using a historical text concerning the outbreak of French Revolution and 5 questions-pairs with alternative answers. We conducted a research with 40 high school students and appropriate historical text and questions in order to have a sample of the distribution of cases to cognitive profiles. The cognitive profiles were judged by hand. Taking into account the experimental results we identified that the frequency of cases with *Very Low*, *Low* or *Nearly Low* cognitive profiles is greater than others. Consequently, as most of the students are expected to have *Very Low*, *Low* or *Nearly Low* cognitive profile, we decided the majority, almost 70%, of initial cases in the case base to be cases belonging to the corresponding subgroups. The system learned the domain of diagnosis of student's HTC from 60 cases: 40 from episodic cases (real students) and 20 prototypical cases. We used weights to indicate which cases merit greater attention as prototypic cases.

4.3 Improving the Scaling Validity

During the preliminary trials, presenting the system with 20 new cases we tested the scaling validity and the performance of the diagnostic system. We divided the test cases into 4 groups of 5 cases each, to assess the improvement in system's performance as it gained experience during the learning process. After each case was entered, the system attempted to assign the correct cognitive profile and profile descriptor to each case. If the model assigned to a case a profile descriptor, which is not correct we added that case into the case base. F-CBR-DHTC, as an incremental knowledge acquisition system, demonstrated satisfactory performance in the four test series.

5 Formative Evaluation of the System

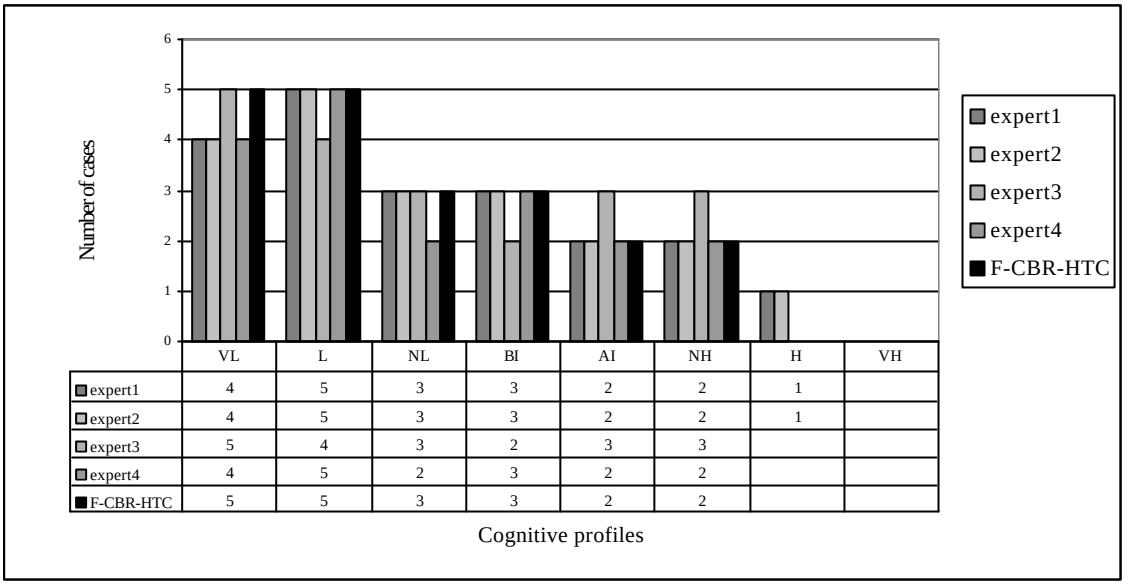
We derived the following questions in order to give structured overview of the used evaluation criteria of the diagnosis system. (1) Does the Fuzzy-CBR system models human behaviour in a more useful way? (2) What is the system's behaviour in novel environment? The evaluation criteria concern noise, user acceptance, flexibility, effectiveness, correctness, consistency and performance.

Formative evaluation aimed at evaluating noise handling in diagnosis accuracy and the system's performance when using a differently structured historical text. 20 High school students, which attend the same public school, and 4 human experts participated in the evaluation. Participation in the experiment was voluntary.

5.1 Evaluation of Noise Handling

In a CBR diagnostic system the problem of *noise*, which means inherent ambiguities, can manifest itself as two or more cases with identical diagnosis but different inputs (Althoff, 1995). In our diagnostic system of students' *cognitive profiles* and *profile descriptors* the diagnosis accuracy has to do with handling the *noise* in occurrences like O: *two or more students with identical cognitive profile have different answers to questions and different profile descriptors*. To account for *noise* within the CBR system's memory, we count the number of occurrences like O and change a certain amount of attributes in the involved cases. During formative evaluation and after the trials, we tested F-CBR-DHTC system for its diagnosis accuracy by 4 human experts in AI. To test our systems' ability to handle noisy data during the adaptation process we conducted a research. The goal of the evaluation research was to examine the extent to which the system and the experts' decisions agree.

In this evaluation stage we presented the system with 25 new (potentially real) cases with slightly different to real cases attributes. After the diagnosis, occurrences like O were found in 20 out of the 25 cases. The system was tested by the human experts for its ability to handle noise. The 4 participants engaged independently in the evaluation and were asked to diagnose the cognitive profiles based on the profile descriptors and judge explaining their view. In most cases there was agreement between the evaluation by the system and the evaluation by the experts.



	VL	L	NL	BI	AI	NH	H	VH
expert1	4	5	3	3	2	2	1	
expert2	4	5	3	3	2	2	1	
expert3	5	4	3	2	3	3		
expert4	4	5	2	3	2	2		
F-CBR-HTC	5	5	3	3	2	2		

Figure 4. Number of cases per corresponding cognitive profile with agreement in the assessment of the profile descriptor using different methods of assessment: 4 human experts and the F-CBR-HTC diagnostic system for 20 cases. The horizontal axis shows the cognitive profiles {Very Low, Low, Nearly Low, Below Intermediate, Above Intermediate, Nearly High, High, Very High}, which correspond to {VL, L, NL, BI, AI, NH, H, VH}.

There were small disagreements in assessing the *cognitive profiles*, when assessment was based on the *profile descriptors*. From the results demonstrated in Figure 4 we can observe that the estimation made by the F-CBR-DHTC diagnostic system and the human experts coincide (on average) nearly in 17 out of the 20 cases. In more details, expert1 coincide in 18 out of 20, expert2 in 19 out of 20, expert3 in 15 out of 20 and expert4 in 17 out of 20 cases.

Regarding the number of students, which were judged as having very low profile, see figure 4, it is remarkable that 3 out of 4 experts judged that one of the 5 very low students is having low profile, nevertheless he has not recognized a cognitive category. To localize the main points of disagreement between the system and the experts, we consider the explanations given by the experts concerning their judgments.

POINT1: The disagreement between experts and the system has to do with considerations of qualitative and not only quantitative characteristics of the profile descriptors as the F-CBR-DHTC system does. For example, an almost complete argument means that the student is about to recognize the corresponding cognitive category. Seen it in terms of learning difficulties, this student is very close to overcome his learning difficulties concerning the cognitive category. So the student is in a higher level than a student with an incomplete argument

POINT2: The disagreement has to do with accidental answers. For example, the student may have accidentally selected the scientific answers for a factor, as it results from the rest of his answers.

Figure 5 depicts analytically the profile descriptors of the five students, which were judged by the system as having very low profile. It also depicts the profile descriptors of the students S1, S2, S3, S4 and S5. Student S5 has given two almost complete arguments, that is the reason why the 3 experts judged him as having low profile.

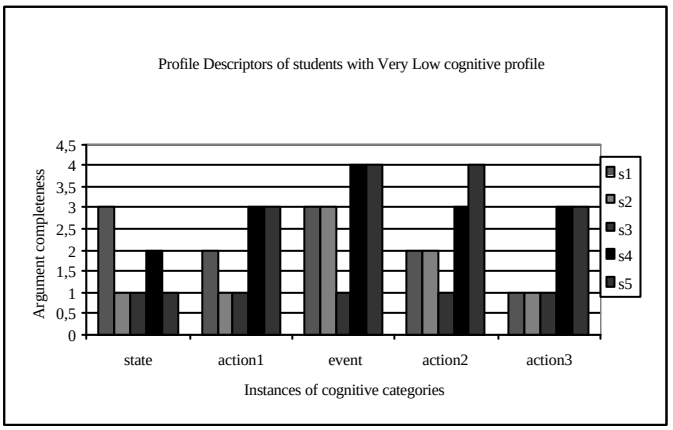


Figure 5. Comparison between the profile descriptors of five students S1, S2, S3, S4 and S5 having Very Low profile according to the F-CBR-DHTC system. The vertical axis shows argument completeness and {1,2,3,4,5} correspond to {incomplete, nearly incomplete, intermediate, almost complete and complete}

Formative evaluation makes us differentiate the students with low profile for example, into two groups: low and low+ according to how close to recognition of cognitive categories they were. Low represents the students that have recognized one cognitive category and low+ represents the students that have recognized one cognitive category but are

very close to recognize another and as a result very close to the nearly low profile, which is cognitive profile of a higher level. The evaluation results were used for the revision of the diagnostic rules taking into account POINT1 and POINT2. After re-initialisation of the case base the revised F-CBR-HTC system becomes more accurate in diagnosis process and can indeed perform diagnosis in a way that gives results similar to the way human experts evaluated students. Consequently the computational system, which imitates human behaviour in a more useful way, obtains user acceptance by human experts.

Table 6
Number and types of recognised cognitive categories and revised cognitive profiles after formative evaluation.

Number of cognitive categories full or closely recognised	Cognitive profiles
no recognition of any cognitive category	very low
close to recognition of one cognitive category	very low+
recognition of one cognitive category	low
close to recognition of more than one instance of a cognitive category	low+
recognition of more than one instance of a cognitive category	nearly low
close to recognition of two cognitive categories nearly	low+
recognition of two cognitive categories	below intermediate
close to recognition of more than two instances of two cognitive categories	below intermediate+
recognition of more than two instances of two cognitive categories	above intermediate
close to recognition of three cognitive categories	above intermediate+
recognition of three cognitive categories	nearly high
close to recognition of more than two instances of three cognitive categories	nearly high+
recognition of more than two instances of three cognitive categories	high
close to recognition of all instances of the three cognitive categories	high+
recognition of all instances of the three cognitive categories	very high

5.2 Evaluation of the system in novel environment

Noise in the data may cause inconsistency in students' behaviour concerning two different historical texts. The evaluation aimed at evaluating the systems' behaviour in novel environment and at validating the system with new data. During the formative evaluation we tested the system's effectiveness when using a new differently structured historical text. We also tested the flexibility and openness of the system, which reflect the environmental dependency and the user acceptability as a feature from the user's interface point of view and time spent to build the application with a new historical text.

The first historical text, which had already been used by the system, concerning the outbreak of French Revolution included 5 factors each corresponding to an instance of a cognitive category. Particularly, it included 3 instances of the cognitive category action, 1 instance of the state and 1 instance of the event. The second was a new for the system historical text concerning the outbreak of the 1st World War (WW). An expert in history teaching prepared the new historical text. The text was consisted of 7 factors each corresponding to an instance of a cognitive category. Particularly, it included 4 instances of the cognitive category action, 2 instances of the state and 1 instance of the event. The expert constructed the 7 question-pairs corresponding to positions and justifications and the appropriate alternative answers and through the interface readjusted the structure of the case base.

20 High school students participated in the two stages experiment. In the first stage the system presented to the students the French Revolution-historical text and they were asked to read the text and select from the alternative answers: (1) their position from a 3point scale stating how strongly each factor they felt was important for the outbreak of French Revolution and (2) the corresponding justification, one out of five proposed, in order to support their position. In the second stage of the research the same students participated and the system presented to the students the 1st WW-historical text and asked the same matters.

At the end, the revised F-CBR-HTC diagnostic system assessed the students' cognitive profiles and profile descriptors for each stage. Comparing the results demonstrated in Figure 5 we can observe that the estimation made by the system in the first stage coincide in 10 out of the 20 students and differs a little in 7 out of 20 students. Even though the sample is rather small to reach a safe conclusion and given the different structure of the two historical texts, the results indicate that the system can give similar assessments for the same student, when performing diagnosis using different historical texts.

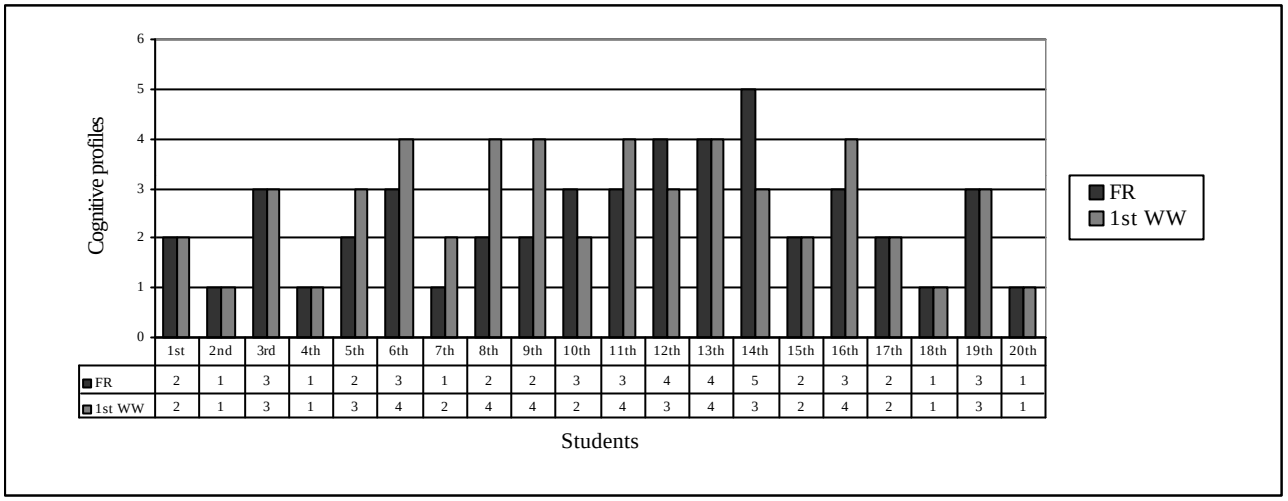


Figure 6. The cognitive profiles assessed by the revised F-CBR-HTC system for 20 students (1st to 20th) using two different historical texts concerning French Revolution and 1st World War: The vertical axis shows the cognitive profiles {Very Low, Very Low+, Low, Low+, Nearly Low, Nearly Low+, Below Intermediate, Below Intermediate+, Above Intermediate, Above Intermediate+, Nearly High, Nearly High+, High, High+ and Very High}, which correspond to {1,2,3,4,5,6,7,8,9,10,11,12,13,14}.

6 Conclusions

In this work, we presented the F-CBR-DHTC system and its formative evaluation, which resulted in the revision of the system. The evaluation focused on the system's diagnosis accuracy of students' cognitive profiles of HTC and on its behaviour in novel environment. The empirical evaluation was preformed with the participation of a small number of experts and on a limited number of students. The results were used for the reformulation of the cognitive profiles in a way that contributes to the improvement of the diagnosis accuracy of the system. The results are encouraging for the system's educational impact on students. The ability of the system to give similar results when using different historical texts holds potential for use in individualized history instruction in an ITS.

Lessons learned from developing and evaluating the system concern practical difficulties due to the uncertainty of knowledge acquisition and knowledge communication among the developer and the human expert. Moreover, worth noticing was the effort required for taking design decisions for the construction of the case base structure, for its enrichment with the appropriate episodic and prototypical cases in order to reassure the performance of the system in the novel environment of a new historical text. The evaluation method adopted was based on the particular characteristics of the system and resulted in the revision and improvement of the system.

The recently growing interest in opening the learner model to the learner encourages the development of systems that give the learner greater responsibility and control over learning. Results of a successful diagnostic process can be beneficial in guiding individualised learning in history by activating the appropriate for a student learning strategy, i.e. through an interactive dialogue between the student and the system. There are educational benefits of the diagnostic system for the students in changing their reasoning. The open learner models are a useful way of helping the experts to recognise learner difficulties, enabling them to respond to specific learner populations or individuals in appropriate ways. We plan to implement the system as an intelligent educational environment.

References

Aamodt A. Knowledge Acquisition and Learning by Experience- The Role of Case-Specific Knowledge, from C. Tecuci, Y. Kodratoff (eds.): Machine Learning and Knowledge Acquisition- Integrated Approaches, Academic press, 1995, 8, 99: 197-24.

Aamodt, A. & Plaza, E. Case-based reasoning: foundational issues, methodological variations, and system approaches. AI Communications 1994, 7(i):39-59.

Althoff K.-D., Evaluating CBR Systems, In: Aadmodt A., Althoff K.-D., Magaldi R. & Mline R. CBR: A New Force In Advanced Systems Development. Tutorial, London, 1995.

Baudet S., Denhière G. (1992). *Lecture, comprehension de texte et science cognitive*, Presses Universitaires de France, Paris.

Bergmann, R. & Althoff, K.-D. Methodology for Building CBR Applications. In: M. Lenz, B. Bartsch-Sporl, H.-D. Burkhard, & S. Wess (Eds.). Case-Based Reasoning Technology from Foundations to Applications, Springer-Verlag, 1998.

Briton B., Graesser Ar., (1996). Models of Understanding Text, Lawrence Erlbaum, Associates Inc. Publishers, Mahwah, New Jersey.

Burke, R. & Kass, A. Refining the Universal Indexing Frame to support retrieval of tutorial stories, Proceedings AAAF94 Workshop on Indexing and Reuse in Multimedia Systems, 1994 , 1-11.

Calmes M., Dubois D., Hullermeier E. et al. A fuzzy set approach to flexible case-based querying: methodology and experimentation, 8th International Conference on Principles of Knowledge Representation and Reasoning (KR2002), Toulouse, France, 2002.

Cavoura Th. Modalités de l' appropriation de la connaissance historique, Thèse de Doctorat, Université de Paris VII, 1994.

Chin D., Empirical Evaluation of User Models and User-Adapted Systems, User Modeling and User-Adapted Interaction, 2001, 11: 181-194.

Dubois, D., Esteva, F., Garcia, P. et al. Fuzzy modelling of case-based reasoning and decision, Case-Based Reasoning Research and Development, Proceedings of the 2nd International Conference on Case-Based Reasoning (ICCBR-97); Leake, D. B., and Plaza, E. (eds.), Springer Verlag, Berlin, 1997, 599–610.

Hansen B. K. Weather Prediction Using Case-Based Reasoning and Fuzzy Set Theory, Phd Thesis, Dalhousie University, Daltech, Halifax, Nova Scotia, 2000.

Jeng, B. C., and Liang, T.-P. Fuzzy indexing and retrieval in case-based systems, Expert Systems With Applications, Elsevier Science Ltd., 1995, 8(1): 135–142.

Kolodner, J. L. Case-Based Reasoning. Morgan Kaufmann, 1993.

Leake, D. B.: CBR in context. The present and future. In: Leake, D. B. (editor) Case-Based Reasoning: Experiences, Lessons & Future Directions, American Association for Artificial Intelligence, Menlo Park California, USA, 1996.

Littman D., Soloway E. Evaluating ITSs: The cognitive science perspective. In Polson M., Ridchardson J., Foundations of ITSs, Lawrence Erlbaum Assosiates Inc. Publishers, Hillsdale, New Jersey, 1988.

Richter M. M., Web S. Similarity, Uncertainty and CBR in PADEX, Automated Reasoning: Essays in honor of Woody Bledsoe (ed. R. S. Boyer), Kluwer Acad. Publ., 1999, 249-265.

Richter, M. M., Burkhard, H.-D. On the Notion of Similarity in Case-Based Reasoning and Fuzzy-Theory, In: Soft Computing in Case-Based Reasoning (eds. S. K. Pal, T. S. Dillon, D. S. Youmd) Springer Verlag, 2001, 29-46.

Schank, R. & Cleary, C., Engines for Education, http://www.ils.northwestern.edu/e-for-e/index.html, 1994.

Shiri A., Aimeur E., Frasson C. SARA: a case-based student modeling system. Advances in Case Based Reasoning 4th European Workshop, EWCBR 98 Proceedings. Springer Verlag, Berlin, Germany, 1998, 394-403.

Tsaganou G., Grigoriadou M., Cavoura Th., Experimental Model for Learners' Cognitive Profiles of Historical Text Comprehension, International Journal of Computational Cognition, 1(4), 2003 (under publication).

Tsinakos A., Margaritis G. Results of employing CBR in SYIM. Learning Technology newsletter, Lov. 3, Issue 4, Oct., 2001.

VanLehn K. Student modeling, In Foundations of Intelligent Tutoring Systems, (eds) Polson M., Richardson J., Lawrence Erlbaum Assosiates Inc. Publishers, Hillsdale, New Jersey, 1988.

Watson, I. Applying case-based reasoning: techniques for enterprise systems. Morgan Kaufmann, California, 1997.

Wenger E. Artificial Intelligence and Tutoring Systems. Computational and Cognitive Approaches to the Communication Knowledge. M. Kaufmann Publishers, Inc., California, 1994.

Weber, G., & Schult, T.J. Case-Based Reasoning for Tutoring and Help Systems. In: M. Lenz, B. Bartsch-Spörl, H.-D. Burkhard, Stefan Wess (Eds.): CBR Technology, From Foundations to Applications, Heidelberg: Springer, 1998.