

Saarland University
Faculty of Natural Sciences and Technology I
Department of Computer Science

Master thesis

**Exercise Metadata Calibration Based on Student Log
Mining**

submitted by

Gautham Adithya Ravi

submitted

11-06-2012

Supervisor

Prof. Dr. Jörg H. Siekmann

Advisor

Dr. Sergey Sosnovsky

Reviewers

Prof. Dr. Jörg H. Siekmann

Priv.-Doz. Dr. Christoph Igel

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement in Lieu of an Oath

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis.

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Saarbrücken,.....

(Datum / Date)

.....

(Unterschrift / Signature)

Acknowledgments

It was a great opportunity to have participated in the 'Intelligent Tutoring Systems' seminar offered by the ActiveMath team just before the spring of 2010. It was here that I sensed artificial intelligence taking an almost human form. I am deeply grateful to my supervisor Prof. Dr. Jörg H. Siekmann and advisor Dr. Sergey Sosnovsky for having provided me with an opportunity to pursue my interest in the field of artificial intelligence in education. I am especially thankful to Dr. Sergey Sosnovsky for the numerous discussions we had on a regular basis towards defining and realizing this work. Without his guidance it would have been very difficult for me to embrace the world of research. Special thanks go to Ahmad Salim Doost, who helped me get started on my work and Eric Andres who proactively helped me avoid many pitfalls that could have been disastrous. Thanks to the ever cheerful and approachable ActiveMath group for the great work environment.

I would further like to thank Rajeev Raman and Saurabh Ray who introduced me to a world of abstract mathematics in the finest way I could imagine. Many thanks go to Ivo Ihrke who demystified for me the then elusive ways of scientific writing. Special thanks go to Kaushik Mukherjee, Mateusz Malinowski, Shyam Sundar and Touseef Liaqat for their great optimism and the encouragement they provided me in exploring the various facets of computer science.

I would like to express my deep gratitude towards my family for the constant support they have provided me and the freedom they have given me in pursuing my interests. Special thanks to my mother for her efforts towards initiating me into cooking, I can't think of a more useful skill I have developed in the recent past. I am more than glad for my friendship with Alexander Horst, Martin Schneider and Vasiliy Melnikov who helped me to a great extent in adjusting to my new student life in Germany.

Abstract

Predicting the parameters of an exercise at the design time is not straight forward. This thesis proposes an approach for re-evaluating learning characteristics of interactive exercises designed for Intelligent Tutoring Systems (ITS). The approach is based on analyzing historical data from the exercise usage logs. It combines retrospective step-by-step prediction of student's knowledge by the user modeling component of an ITS with the psychometric apparatus of Item Response Theory (IRT). The algorithm tries to take into account the reliability of log data and estimates the overall degree of trust for every suggested calibration. The evaluation of the proposed approach has demonstrated high predictive validity of the newly computed metadata.

Contents

Chapter 1 Introduction	1
1.1 Intelligent Tutoring Systems	1
1.2 Learning Object Metadata	3
1.3 Motivation and Approach.....	4
Chapter 2 Related Research	7
2.1 Introduction to Educational Data Mining.....	7
2.2 EDM methods taxonomy	8
2.3 EDM for Student Modeling & Performance Prediction	10
2.4 EDM for Estimating Exercise Characteristics	12
2.5 Relevance to the Proposed Approach.....	13
Chapter 3 Background Theory	15
3.1 Item Response Theory	15
3.2 Knowledge Tracing.....	17
Chapter 4 Approach	21
4.1 Log Filtering and Enrichment.....	22
4.2 Estimation of Student's Ability	23
4.3 Calibration of Exercise Metadata.....	25
4.3.1 Clustering Exercise Events	25
4.3.2 Computing Probabilities of Successfully Solving an Exercise	25
4.3.3 Estimating the Exercise Parameters.....	26
Chapter 5 Evaluation	27
5.1 Datasets	27
5.1.1 Hochschule für Technik und Wirtschaft (HTW) Learning Content	28
5.1.2 Adaptive Tutorial Feedback (AtuF) Project.....	29
5.2 Difficulty Estimation	30
5.3 Testing Calibrated Difficulty.....	33
5.3.1 HTW test set predictions.....	35
5.3.2 AtuF test set predictions	36
5.4 Discrimination Estimation	38

Chapter 6 Conclusions and Discussion	41
6.1 Evaluation Design.....	41
6.2 Differences in Evaluation Data	42
6.3 Approach Comparison	43
6.4 Future Work.....	45
References.....	47

Chapter 1

Introduction

Nowadays, technology-based solutions are widely adopted to facilitate and improve learning all over the world. E-learning has emerged as a convenient and cost effective paradigm of learning with the widespread implementation of computer networks. It has especially been successful in scaling up the learner group size and diversity by extending the reach of education beyond the physical constraints of a classroom. In addition, the learner gains convenience and flexibility as a result of the independence from fixed schedules and venues. However, one important aspect that most e-learning systems usually fail to address in a satisfactory way is the incorporation of individualized tutoring, which is widely believed to be the most effective method of teaching (Bloom, 1984). For e-learning to be maximally effective, it needs to capitalize on the learner characteristics and deliver the learning content in an adaptive manner. A potential solution involves development of intelligent tutoring systems that seek to replicate one-on-one tutoring by personalizing the learning experience of every individual learner, using tools and techniques predominantly from the fields of artificial intelligence, data mining and human-computer interaction.

1.1 Intelligent Tutoring Systems

Several decades of developing intelligent tutoring systems (ITS) have generalized its architectural framework into a four component model (Nkambou et al., 2010) as illustrated in Figure 1.1.

- The domain model (also called expert knowledge) that stores and structures all the concepts, rules and/or problem-solving strategies of the learning domain. It can be perceived as a representation of the domain expert's knowledge or a frame of reference

for evaluating a student's learning performance to detect misconceptions and gaps in student knowledge.

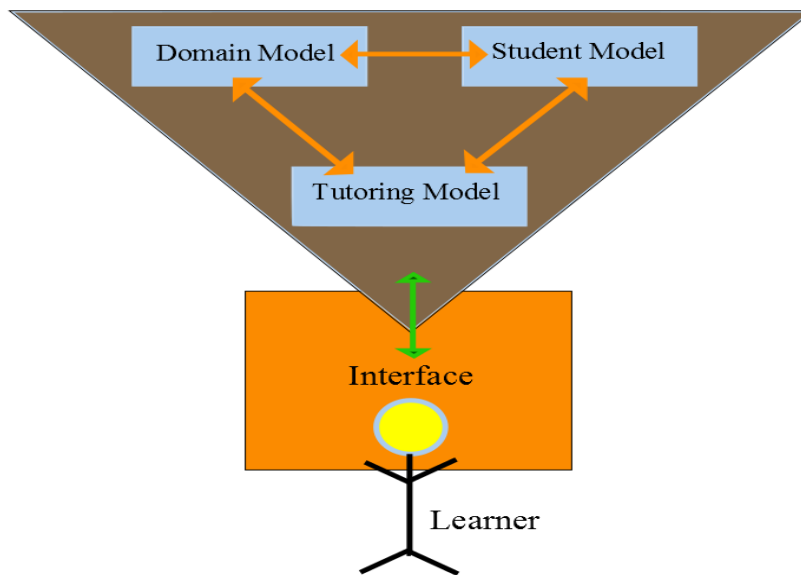


Figure 1.1: ITS - generalized architecture

- The student model that registers all available information on the relevant characteristics of a student (such as knowledge, skills, background etc.) using which it infers on the possible changes in the student's knowledge and learning processes. This component plays the key role in personalization of the tutoring experience since without it all students would be treated similarly.
- The tutoring model (also called pedagogical or adaptation model) formalizes the pedagogic expertise for guiding the adaptation process based on the information made available by the domain model and the student model. Its functionality also includes enhancing the learning experience effectively by means of personalized interaction with the learner (eg. timely interventions in the presentation of learning content with Socratic dialogues, hints, feedback from the system etc.).

- The interface component enables the student's access to the learning material. It is typically in the form of a graphical user interface (GUI) and plays a very important role in sustaining learner interaction which is the key to intelligent tutoring.

1.2 Learning Object Metadata

Learning objects in an e-learning environment are the elementary parts used for constructing instructional material. They can represent exercises, examples, definitions, quizzes, etc. These learning objects are annotated with metadata to convey information on their pedagogical aspects and enable indexing in learning content repositories. The learning object metadata support not just simple retrieval and versioning of learning content, but enables tracing and interpretation of students' knowledge and reasoning about learning objects in terms of their characteristics and the elements of the domain semantics that they cover.

Educational metadata, as illustrated in Figure 1.2, can be classified as descriptive, semantic or pedagogic. The descriptive metadata (such as title, date, language, creator, rights etc.) is reliably annotated by the author during content design, but estimation of the semantic metadata (typically involving domain dependent, fine grained, structural and semantic information annotated by a domain expert) and pedagogic metadata (typically involving information on how the object would function on a learner, such as difficulty, competency, typical learning time etc) is not as simple.

Some pedagogic metadata can be statistically estimated by observing functioning of a learning object on an appropriate and sizeable group of learners. Such observations can be especially informative when the learning object is interaction oriented and involves some means of direct feedback from the learner e.g. an exercise or a quiz. In the context of an ITS, one of the most important elements of the pedagogic metadata is the learning object difficulty level,

which can be estimated empirically and used to compliment and/or validate the annotation accuracy of a human expert.

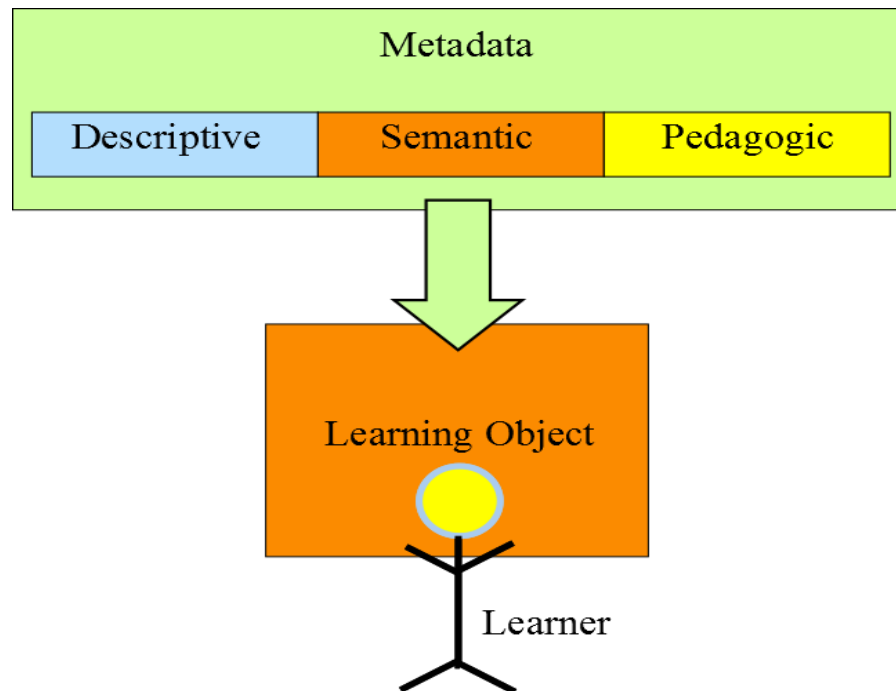


Figure 1.2: Learning content metadata

1.3 Motivation and Approach

The importance of accurate learning content metadata in an e-learning environment cannot be overstated, since it is the key to the aspects of reusability, adaptability and extensibility of e-learning content. Missing, incomplete and poorly constructed metadata is one of the major issues affecting the functionality and efficiency of e-learning systems. For instance, inaccurate calibration of the item difficulty can result in the failure of learning in an ITS in the following ways.

- Presenting easy learning objects that are incorrectly annotated as difficult, to a student very proficient in the underlying concepts, will result in boredom.

- Presenting difficult learning objects that are incorrectly annotated as easy, to a student not proficient in the underlying concepts, will result in a sequence of failures and frustration.

This thesis aims at accurate calibration of the pedagogic metadata of exercises in an e-learning system by relying on a Bayesian framework of student knowledge estimation for every exercise response registered. The proposed approach, as shown in Figure 1.3, is based on mining students exercise activity logs to estimate students' proficiency of underlying concepts. These estimations are then used to infer the pedagogic characteristics of exercises used in the course. The approach is then evaluated based on the predictive performance of the calibrations.

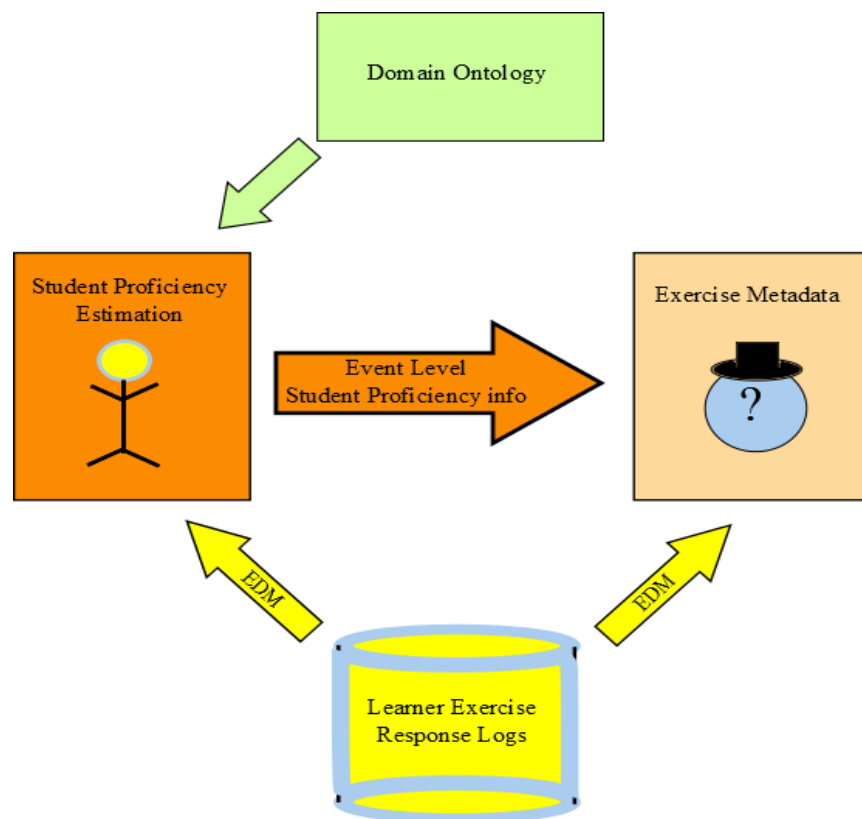


Figure 1.3: Generalized approach

The thesis is partitioned into 6 chapters, with Chapter 2 surveying the related research on similar lines, Chapter 3 introduces the reader to important background theories involving the modeling of exercise response and student ability that facilitates a concrete understanding of the work in this thesis, Chapter 4 illustrating the approach used in reaching the objective, Chapter 5 dealing with the evaluation and visualization of the results and Chapter 6 summarizing the work and listing indicators and possibilities for further research.

Chapter 2

Related Research

Data mining is the field of research that aims at extracting hidden patterns from existing data, which could reveal unknown information about the data and/or the data generation process. Mining of student activity logs in e-learning systems enables improving and validating our understanding of students and the settings they learn in. Due to the multi-level hierarchy and non-independent nature of educational data, it has been proposed that educational data mining deviates from conventional data mining methods towards psychometric modeling frameworks, which have shown to be effective in the analysis of the learning process (Baker & Yacef, 2009). This section serves to introduce the reader to the field of Educational Data Mining (EDM) and lists research in this field that relates most to the direction of the work presented in this thesis.

2.1 Introduction to Educational Data Mining

EDM, considered today as an emerging discipline, focuses on developing methods for exploring the characteristics of the data involved in educational contexts. Typically relying on the historical log data of all possible student activity in the system, EDM-based techniques have contributed to enhancing e-learning especially in the areas of

- Course report generation for analysis and visualization
- Student modeling and performance prediction
- Maintenance and improvement of learning content
- Domain structure analysis
- Recommendation generation

2.2 EDM methods taxonomy

Romero and Ventura (2007), with a web data based perspective of student activity logs have categorized the work in EDM into the following categories:

- Statistics and visualization
- Web mining
 - Clustering, classification, and outlier detection
 - Association rule mining and sequential pattern mining
 - Text mining

A more recent categorization by Baker (2010) considers a more objective perspective.

- Prediction
 - Classification
 - Regression
 - Density estimation
- Clustering
- Relationship mining
 - Association rule mining
 - Correlation mining
 - Sequential pattern mining
 - Causal data mining
- Distillation of data for human judgment
- Discovery with models

Prediction aims at inferring a single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables) and requires having labels or a scale for its accuracy to be determined. Prediction techniques in EDM correspond to a fair degree to the techniques used in data mining, though there is lesser emphasis on popular

techniques such as support vector machines and neural networks, and relatively more emphasis is noticeable towards linear methods. Detection of student behavior, affect and emotion has been effectively utilized in predicting the differences in student learning and harnessing it to improve the adaptability of an ITS (Walonoski & Heffernan 2006).

Clustering deals with appropriate grouping or categorizing of data instances based on the similarity in the nature of the data rather than on any pre-assigned labels. Clustering techniques used in EDM have been based on popular clustering methods such as k-means, Expectation Maximization (EM) and Gaussian Mixture Model based frameworks. Student models in ITS can be trained based on the clustering of student interactional behavior (Amershi & Conati 2009).

Relationship mining techniques have been more prominent in aiding researchers and end users of ITS in the data analysis and inference processes. Relationship mining has also been used for capturing and improving the modeling of the domain structure where the domain knowledge is not predetermined or is not well defined (Nkambou et al. 2007).

Distillation of data for human judgment involves representation and visualization of data for human inference and helps humans solve problems and/or make decisions that are not supported by the automation offered by existing data mining techniques. This can be especially useful in understanding student behavior. Important factors affecting the learning process, such as learner motivation, can be assessed based on diagnosis and heuristics (HersHKovitz & Nachmias 2008).

Model based discovery aims at developing models of phenomenon involved in e-learning based on processes that can be validated in some way and using them effectively as components in other analysis. It has been used for supporting sophisticated analyses especially in the context of student behavior such as impact of different types of student behavior on learning style (Cocea et al., 2009) and effect of ITS design variations on learner behavior (Jeong & Biswas, 2008).

Inference on the learning object characteristics using EDM techniques ranges from association rule mining and clustering of learning content (Zaiane, 2002; Tang & McCalla, 2005), more dominant during the earlier years of EDM to prediction performance, more recently enabled by the increasingly available public data and the integration EDM has experienced from the field of student modeling (Barnes, 2005 and Desmarais & Pu, 2005). A complete review of all the work corresponding to EDM is beyond the scope of this thesis, so in the next subsections focus is put upon the two most important aspects that enable adaptability in e-learning and are the main focus of this thesis, the student cognition modeling and the exercise characteristics (Figure 2.1).

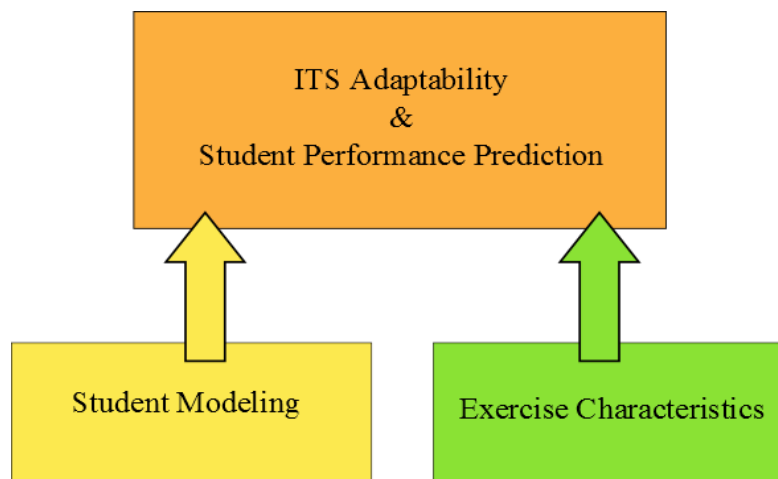


Figure 2.1: Thesis – Areas of focus

2.3 EDM for Student Modeling & Performance Prediction

The primary aim of student modeling in e-learning is assessing the degree to which a student has learnt and tracking when and how this learning occurs. One major contribution of EDM methods in the recent years has been towards introducing several new ways to enhancing data driven student modeling approaches, with a direct relevance to the predictive accuracy of student performance. With the introduction of the PSLC's DataShop (Koedinger et al., 2010), an open repository of e-learning data, predicting student performance (PSP) was the chosen

challenge at the KDD Cup 2010. Typically classification and regression techniques (based on statistics, decision trees, rule induction etc.) have been used for performance prediction when the objective of prediction is clear (e.g. predicting the correctness of a students' response to an exercise, speed of student response etc) and clustering techniques, resorted to in places of ambiguity (such as detecting different kinds of student behavior).

Research on data driven methods in automating the construction of cognitive process models based on solution path mining and production rules has existed as early as the 1980's (e.g. Sleeman, Langley, & Mitchell, 1982; Langley & Ohlsson, 1984). By the mid 1990's, student models that inferred the probability of a student knowing a specific skill or concept based on the student performance data had been introduced (cf. Corbett & Anderson, 1995; Martin & VanLehn, 1995; Shute, 1995). Since then there have been significant developments in validating the precision of student models, recognizing phenomena that interfere with a regular learning process (such as following erroneous strategies, gaming the system (Baker et al. 2008) & off-track student behavior (Baker, 2007)) and modeling of student meta-cognitive skills (e.g. help-seeking and self-explanation) and affective states (such as motivation and emotion).

Different methods have been used for student modeling. Rus et al. (2009) compare six different algorithms (Naïve Bayes, Bayes Nets, Support Vector Machines, Logistic Regression and two variants of decision trees) in detecting student mental models using student generated paragraphs during prior knowledge activation and find Bayes Nets to be the most accurate. The dominant influence of student modeling on the predictive performance of EDM based approaches initially enabled by modeling processes such as Knowledge Transfer (Corbett & Anderson, 1995), continues to develop with techniques such as Learning Factor Analysis (Cen et al., 2006) and Matrix Factorization (Thai-Nghe et al., 2011). Q-matrix method (Barnes, 2005) uses EDM to create concept models for the learning content and uses them to understand student behavior and direct learning paths for future students. Rai et al. (2009) propose using Dirichlet priors in developing more plausible models of student knowledge. Recently,

SimStudent an intelligent agent based on production system rule learning for automatic student model discovery has been evaluated to perform better than human generated knowledge component models (Li et al., 2011).

2.4 EDM for Estimating Exercise Characteristics

Less work has been done in the field of EDM on estimating learning content characteristics, compared to EDM-based student modeling, though the quality of adaptation depends on the quality of content models as well. A study by Gong & Beck (2011) suggests that exercise difficulty is a better predictor of student performance than skill difficulty or the modeled student proficiency. Gong et al. (2010) have compared Performance Factor Analysis (PFA) (Pavlik et al., 2009), which predicts student performance based on exercise difficulty and student logs with Knowledge Tracing (KT) (Corbett & Anderson, 1995) based approaches that do not use exercise difficulty, and though the study didn't show any real difference in their predictive accuracies, the authors hypothesize that PFA works better in circumstances where exercises for a particular skill/concept vary greatly in difficulty.

The notion of exercise difficulty has been rigorously studied under Item Response Theory (IRT) (van der Linden & Hambleton, 1997), a technique effectively used for assessing students in computer adaptive testing (CAT) (EDS, n.d.). Wauters et al. (2010) explored the feasibility of applying IRT in adaptive item-based ITS and list two main challenges, the dataset and the algorithm. Exploring alternative means of estimating the exercise difficulty, Wauters et al. (2011) find that the performance of a simple proportion correct based exercise difficulty measure correlated similarly to IRT with the true difficulty followed by student feedback, Elo rating, expert rating, paired comparison (student) and paired comparison (expert), though they claim to have used a rather small sample size. Kernel smoothing based exercise calibration (Guzmán & Conejo, 2005) for adaptation using polytomous response models has been proposed to be

effective in comparison to conventional IRT based methods for a reduced number of prior student sessions and has been implemented in the SIETTE system (Conejo et al., 2004).

Categorical and sequential characteristics of exercises have been studied as well. An attempt to predict the group of exercises that induced learning using hypothesis on the ASSISTments system (Razzaq et al., 2005) was unable to account for the irregularities in learning across the groups of skills (Feng et al., 2008). The significance of order in exercise sequences has been simulated and shown to hold potential in optimizing learning (Pardos & Heffernana, 2009).

2.5 Relevance to the Proposed Approach

The work in this thesis focuses on accurate data-driven calibration of exercise difficulty metadata to improve the accuracy of student performance predictions. This is done by applying clustering and regression techniques to student proficiency information obtained using a standard Bayesian Knowledge Tracing (Corbett & Anderson, 1995) based student model and the outcomes of student responses to e-learning exercises. The approach is then evaluated by predicting the performance of future students on the calibrated exercises and comparing it with the results of their actual performance.

The calibration of the exercise difficulty follows the Item Response Theory model (IRT) (van der Linden & Hambleton, 1997). Both, the student modeling scheme and the exercise metadata calibration scheme that are used in the proposed approach are empowered by EDM.

Chapter 3

Background Theory

This section introduces 2 modeling techniques that the proposed approach is built on. IRT is used for modeling the student's activity with an exercise and KT, for modeling student's knowledge of a concept.

3.1 Item Response Theory

Item Response Theory (IRT) helps to estimate the probability of a correct response to a test item given the ability of an examinee (Lord, 1980). Thus, it provides a mapping model between the student's parameters and the parameters of an exercise (such as difficulty) by bringing them to the same scale. The model is expressed as a logistic (sigmoid) function. Equation 3.1 represents a 2-parameter item model:

$$Pr(\theta) = \frac{1}{1 + e^{-discrimination(\theta - difficulty)}} \quad (3.1)$$

The probability $Pr(\theta)$ of a student with the ability θ to answer the item correctly depends on the two parameters of the function:

1. The point on the ability scale where the probability of answering the item correctly $Pr(\theta)$ is equal to 0.5 represents the difficulty of an item. Difficulty defines the required ability level, at which it becomes more probable to answer the item correctly.
2. The rate, at which the logistic function grows, translates to the measure of the item's discrimination between the students who can successfully solve the item and those who cannot. The higher the value of the function slope, the better the test item can distinguish between the students who have the necessary ability level to solve the exercise successfully and those that do not.

In addition to the 2-parameter model, there exist also the 1-parameter model that treats all test items as having equivalent discriminations and 3-parameter model that assumes the possibility of guessing.

Figure 3.1 presents an example of the 2-parameter model for a test item with the discrimination = 15 and difficulty = 0.65. It uses a modified ability scale. IRT provides little guidance on the exact meaning of ability, and how it should be measured and interpreted. Traditionally, the ability axis uses a scale from -3 to 3; individual values on this scale do not have a definitive meaning, besides, when compared to each other. In this work, ability is interpreted as a probability that a student has mastered the concept required to answer the exercise. Consequently the scale for students' ability ranges from 0 to 1 (see Figure 3.1).

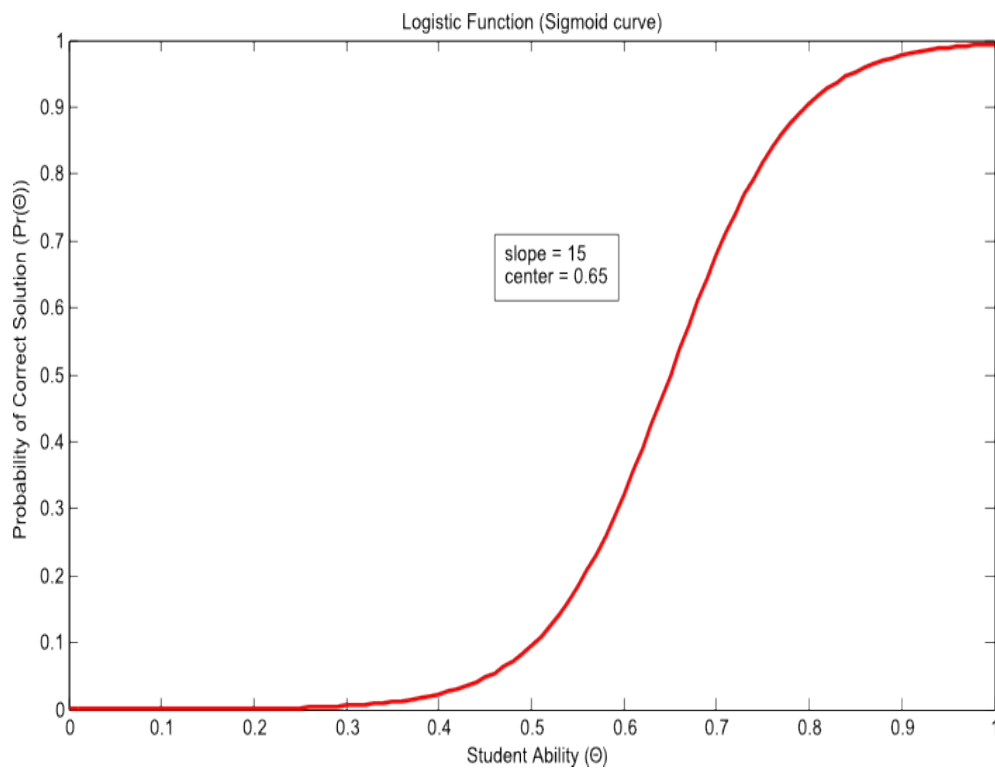


Figure 3.1: Item response function (sigmoid curve)

The approach used in this thesis uses a 2-parameter IRT model to estimate the exercise difficulty and discrimination metadata from the mined information on the student abilities and their probability of answering the exercise correctly.

3.2 Knowledge Tracing

Knowledge Tracing (KT) is a popular algorithm for probabilistic estimation of student's mastery of knowledge components (skills and concepts) based on the history of attempts to solve related exercises (Corbett & Anderson, 1995). Given that a student's attempt to solve an exercise can be classified as either correct or incorrect and considering that each exercise is associated with a knowledge component, a four parameter KT algorithm models the evolution of student's proficiency in applying these knowledge components based on the following rules of inference (Beck & Sison, 2006).

- Each item addresses only one knowledge component (extensions available for items handling multiple knowledge components, but we stick to the naïve model).
- The latent trait of a student for a knowledge component can either be in a learnt state or in a not-yet-learnt state at any given moment.
- There exists a finite probability that a student learns a knowledge component while attempting to solve an item addressing it, but the case of unlearning a knowledge component doesn't exist.
- There exists a finite probability that a student solving an item correctly actually guessed, i.e. knowledge component of the item for the student is not in a learnt state yet even though the item was solved correctly.
- There exists a finite probability that a student solving an item incorrectly actually slipped, i.e. knowledge component of the item for the student is in a learnt state even though the item was not solved correctly.

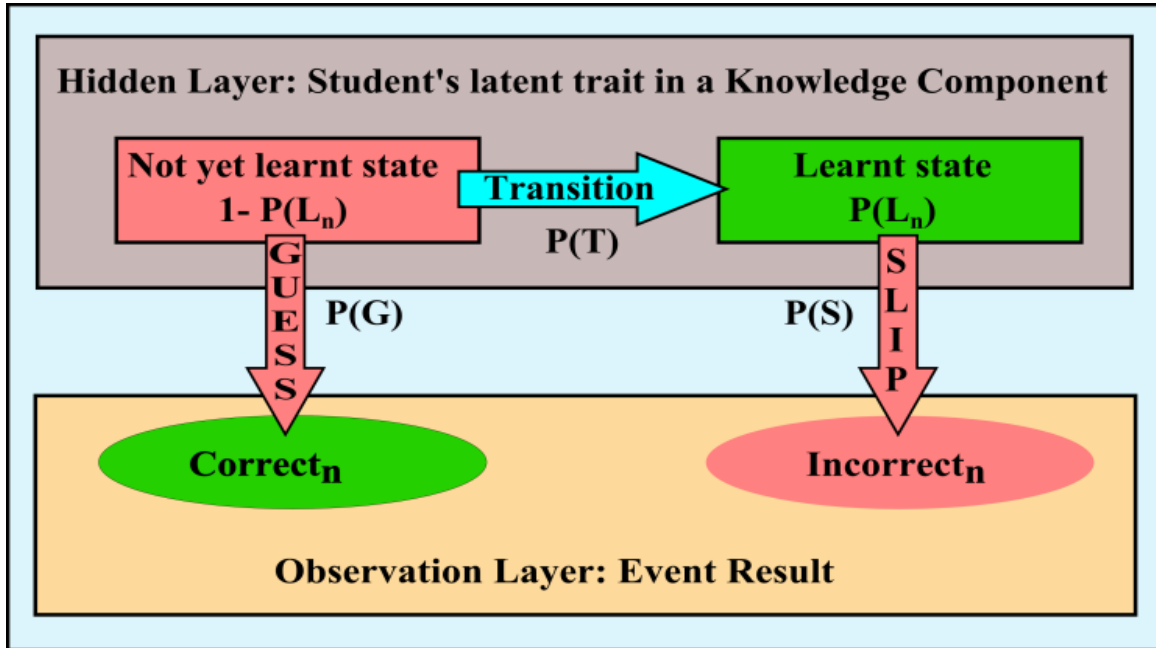


Figure 3.2: Knowledge Tracing (4-Parameter Model)

Figure 3.2 illustrates the basic mechanism of a four parameter KT model. After every attempt to solve an exercise (i.e. to apply a corresponding knowledge component), the conditional probability of the student having mastered the knowledge component $P(L_n | Result_n)$ is re-computed based on the following four parameters as described in Table 3.1 and equations 3.2, 3.3 and 3.4 (Corbett & Anderson, 1995).

Table 3.1: Parameters of KT algorithm (4 parameter model)

$P(L_{n-1})$	Probability that the knowledge component has been in a learnt state before the attempt.
$P(G)$	Probability that a correct attempt was a result of a <i>guess</i> .
$P(S)$	Probability that an incorrect attempt was caused by a <i>slip</i> .
$P(T)$	Probability that a student learns a knowledge component while attempting to solve an item addressing it.

$$P(L_{n-1}|Correct_n) = \frac{P(L_{n-1}) * (1 - P(S))}{P(L_{n-1}) * (1 - P(S)) + (1 - P(L_{n-1})) * P(G)} \quad (3.2)$$

$$P(L_{n-1}|Incorrect_n) = \frac{P(L_{n-1}) * P(S)}{P(L_{n-1}) * P(S) + (1 - P(L_{n-1})) * (1 - P(G))} \quad (3.3)$$

$$P(L_n|Result_n) = P(L_{n-1}|Result_n) + (1 - P(L_{n-1}|Result_n)) * P(T) \quad (3.4)$$

The main approach presented in this thesis, does not have any constraints regarding the choice of student modeling technologies. We use KT, because it is a well-established algorithm for estimating student's knowledge from their training history, and this allows us to assume a certain level of universality of the obtained results.

Chapter 4

Approach

An ITS typically maintains a log of all activities performed on it. The logged activity data could be stored in various formats, but generally include basic information on the activity, such as the time and location of an event, the actor involved, the object involved, the action performed and information on the consequences of the action. When considering exercises as our objects, we have students as our actors and their actions defined by their solution provided for the exercise. The consequences of the action of the student on an exercise in its simplest form is the correctness of the solution provided by the student, which is measured by the domain model of the ITS in the appropriate way. The domain model of the ITS also provides the mapping between the various knowledge components involved in the learning content of the ITS and their manifestation in the form of exercises. This mapping is usually extracted from ontology of the domain model for the learning content.

The proposed approach aims to calibrate the exercise difficulty and discrimination metadata using the basic event information from the logged activity data on exercises and the mapping between the exercises and knowledge components. At a high level, the proposed approach consists of 3 main phases.

1. Log filtering and enrichment;
2. Estimation of student's ability;
3. Calibration of exercise metadata.

Figure 4.1 graphically presents the overview of the approach with the transition from one phase to another. Next three subsections describe the details of every phase.

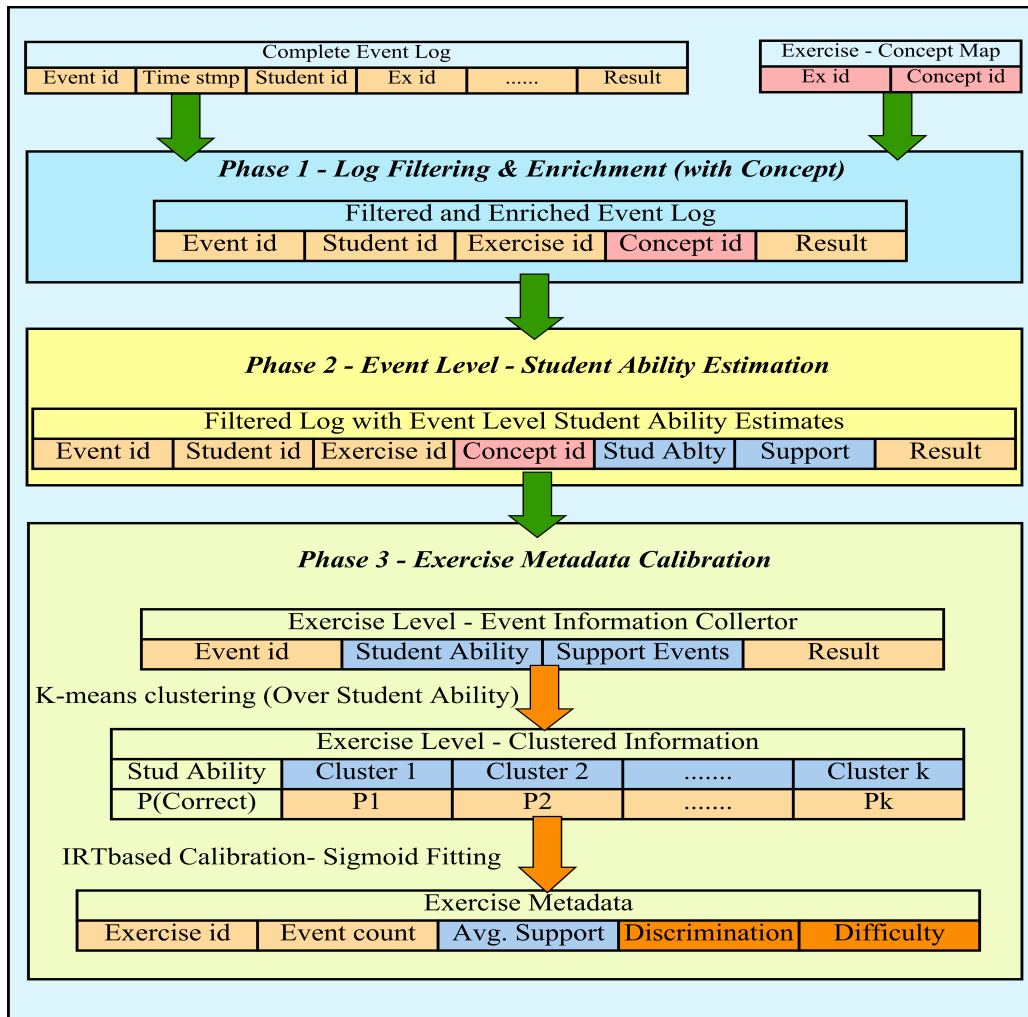


Figure 4.1: Approach overview

4.1 Log Filtering and Enrichment

Before the logs can be processed by the main algorithm, they have to be cleaned to contain only the relevant events. For this purpose the unique set of ITS users and exercises involved in the log data are extracted. The log data is then filtered to contain only events involving students, ignoring any irrelevant activity such as events generated by the authors of the learning content or administrators of the ITS. The log data is also filtered to exclude any exercises that had missing or improper annotation of knowledge component mapping. The student's activity with an ITS that incrementally generates the student-log usually contains a wide range of events such as 'User login/out', 'Page requested/presented', 'Note created/modified/deleted', 'Tool

started/exit' etc along with the 'Exercise step/finish' events where the student attempts to solve an exercise. The student-log data is then filtered to contain only the events of direct relevance in the modeling of student – exercise interaction i.e. the 'Exercise step/finish' events which give us concrete information on the result of the activity involving the students' ability. The resulting log data obtained after filtering out the irrelevant ITS users, exercises and events can then be considered clean.

The event count registered for each student and each exercise in our clean log data is obtained and the cleaned log data is further filtered to contain only students and exercises with more than 10 events in total. This is done because the ability estimates of students with low interaction history and calibration of exercises with little data to support the calibration process is unreliable. The minimum threshold for the student is decided based on the average student event count and for the exercise is decided based on the number of different ability levels the student group are considered to have. We then finally enrich our completely filtered log data with information on the underlying knowledge component (concept) addressed by the exercise in each event.

4.2 Estimation of Student's Ability

The filtered log data is then further enrich with estimates of the student probability of having mastered the underlying concept of each event based on the KT model. This is done by updating the current conditional probability of the student having mastered the underlying concept based on the probability of mastery from the previous interaction between the student and concept, given the result of the current interaction. For each event the number of previous events corresponding to the same student and concept is also registered, since this serves as a measure of the support the current estimation has from the past observations.

This phase, as illustrated in Figure 4.2, is implemented by maintaining a separate student-concept matrix. The student-concept matrix for the training set students is initialized

with the initial probability of a student having mastered the concept. Iterating over the events of the filtered log data, the student-concept matrix is updated after each event based on the KT model and the updated probability of mastery along with the number of times the probability of mastery for the particular student-concept pair has been updated (support) is spontaneously appended to the filtered log data.

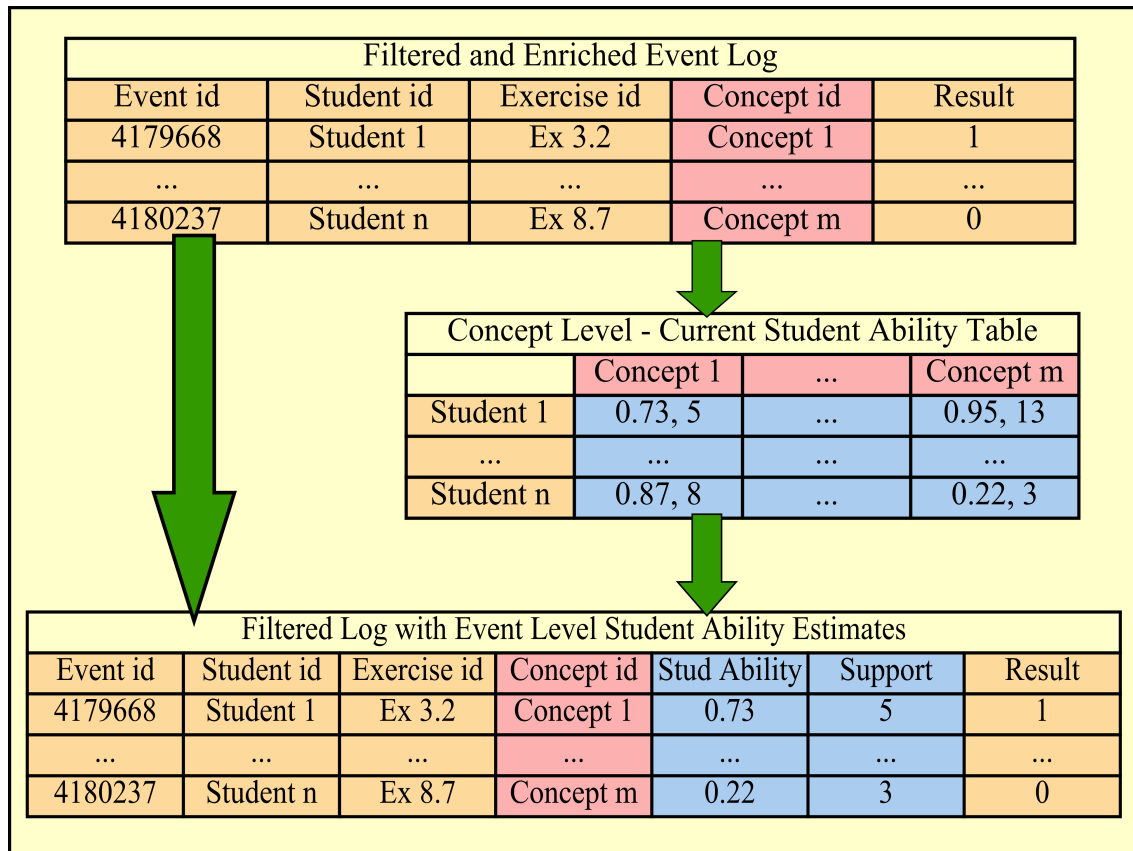


Figure 4.2: Event level – student ability estimation

It is important to note that the student abilities corresponding to each event are estimated following the chronological order of the events in the log data and hence the sequential information of the events is also utilized in estimation process. The filtered log data now includes information about the underlying concept, the probability of the student having mastery of the concept for each event and the count of the supporting events for the estimation of the probability of mastery (support).

4.3 Calibration of Exercise Metadata

The final phase of the proposed approach computes the IRT based estimation of the exercise difficulty and discrimination metadata. Figure 4.3 illustrates this phase, where the optimal parameters of the sigmoid curve that best describe the observations for an exercise are estimated.

4.3.1 Clustering Exercise Events

An array for each exercise in our training set log is created and events corresponding to each exercise are appended to their corresponding arrays while iterating over the log data. For every event we append 'n-1' replicas of the event to the exercise array, where 'n' represents the floor of the natural logarithm of the support of the event. This is done to put a greater emphasis on the events that have a greater number of supporting events and can be seen as a logarithmic weighing of the significance of the event based on the support it registers. The events for each exercise are then clustered using a standard k-means clustering algorithm (MacQueen, 1967) into 10 ability groups based on the student's probability of mastery in the concept for the event. The k-means clustering algorithm minimizes the aggregate distance between each event's student ability estimate and the mean of the ability group the event comes under, by iteratively optimizing the mean values of the ability groups, given the desired number of ability groups.

4.3.2 Computing Probabilities of Successfully Solving an Exercise

For each exercise the probabilities for the different ability groups to solve the exercise correctly are then computed by taking a ratio of the number of events with a correct solution to the total number of events, corresponding to each ability group registered for the exercise. The mean ability level of each ability group is considered as the indicator for the group's ability level, gives us 10 pairs of ability values and probability of the ability value solving the exercise correctly. This

gives us the required observations for estimating the exercise difficulty and discrimination parameters using the 2 parameter logistic (sigmoid) function.

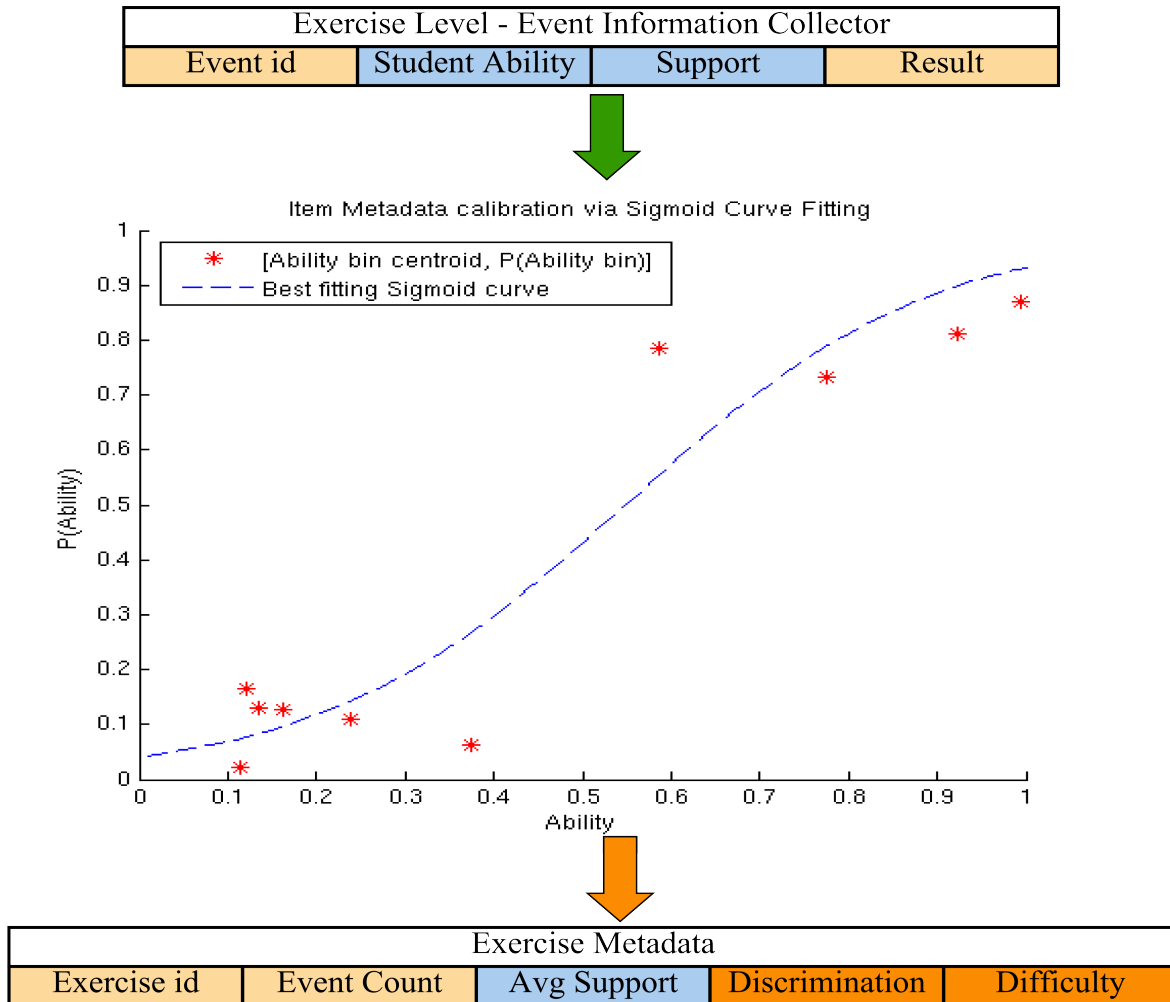


Figure 4.3: Exercise metadata calibration

4.3.3 Estimating the Exercise Parameters

For each exercise a sigmoid curve is then used to represent the 10 ability - probability (ability) observation pairs by minimizing the least squares distance between the curve and the observation pairs using the trust-region-reflective algorithm (Coleman & Li, 1996). The parameters of the optimal sigmoid curve that fits the observations correspond to the difficulty and discrimination estimations of the exercise.

Chapter 5

Evaluation

Evaluation was performed based on splitting the students in our dataset into training set, containing 80% of our students and test set, containing the other 20%. The difficulty and discrimination metadata for the exercises used in our dataset are estimated using the events involving the training set students. For each event involving a test set student a correct response is predicted if the student ability is greater than or equal to the exercise difficulty. Comparing the predictions with the actual results, the predictive values, predictive rates and F-1 measure are then measured. This translates to the evaluation of the estimated difficulty metadata of the exercises.

The evaluation of the discrimination metadata estimates is not as straightforward as the evaluation of the difficulty metadata estimates, because of the unavailability of an absolute measure that it could be compared against. Hence, only a summary of the estimated discrimination metadata is provided and usage of the discrimination metadata in student performance prediction and learning content analysis is discussed in future work section of chapter 6.

5.1 Datasets

The evaluation was performed on the student activity log generated by the ActiveMath e-learning system (ActiveMath, n.d.), on content developed for the Hochschule für Technik und Wirtschaft (University of Applied Sciences), Saarbrücken (HTW, n.d.) and the Adaptive tutorial Feedback project (AtuF, n.d.).

5.1.1 Hochschule für Technik und Wirtschaft (HTW) Learning Content

The HTW learning content used in the evaluation consisted of training exercises designed towards providing a bridging course to cover the knowledge gaps in high school mathematics for students educated in the different schooling systems prevalent in Germany. The exercise activity was recorded over a period of 1 month in September – October 2009, with a non-specific student group working with 95 different exercises covering a total of 33 concepts involving high school level mathematics. An example of the exercises in the HTW learning content is shown in Figure 5.1.

ActiveMath
Drucken | Hilfe | Impressum

Übung

1. Sei A eine Menge. Geben Sie $A \cap A$ und $A \cup A$ an!

Füllen Sie die Felder aus!

$A \cap A =$

$A \cup A =$

2. Es gilt $A \subseteq B$. Geben Sie $A \cap B$ und $A \cup B$ an!

Füllen Sie die Felder aus!

$A \cap B =$

$A \cup B =$

3. Welche Eigenschaften der Vereinigung und des Durchschnittes von Mengen gelten? Wählen Sie das richtige Zeichen aus!

$A \cap (B \cup C) = (A \cap B) \mid \left[\begin{array}{c} \subseteq \\ \supseteq \end{array} \right] \mid (A \cap C)$

und

$A \cup (B \cap C) = (A \mid \left[\begin{array}{c} \subseteq \\ \supseteq \end{array} \right] B) \mid \left[\begin{array}{c} \subseteq \\ \supseteq \end{array} \right] \mid (A \mid \left[\begin{array}{c} \subseteq \\ \supseteq \end{array} \right] C)$

Input Editor aktivieren

Figure 5.1: Typical exercise in the HTW content delivered by the ActiveMath system

A summary of the HTW evaluation data is listed in Table 5.1.

Table 5.1 Evaluation dataset details (HTW data)

	Total	Training set	Test set
Events	13234	10741	2493
Students	371	298	73

5.1.2 Adaptive Tutorial Feedback (AtuF) Project

The AtuF project involved a controlled study to investigate the effect of self-adaptive feed-back in the domain of fraction arithmetic. The study involved a pre-test to detect student misconceptions, a treatment phase to correct the misconceptions and a post-test to evaluate the effectiveness of the treatment phase. Only the post-test exercises were used in the evaluation of our approach. The log data was recorded over a period of 6 weeks in February-March 2009, with the participation of 6th and 7th grade students from schools in Dresden. A total of 24 different exercises covering 5 different concepts from fraction arithmetic were used. An example of the exercises in the HTW learning content is shown in Figure 5.2. A summary of the AtuF evaluation data is listed in Table 5.2.

Table 5.2 Evaluation dataset details (AtuF data)

	Total	Training set	Test set
Events	2751	2250	501
Students	186	149	37

ActiveMath – Übung

141.30.40.34:8080/ActiveMath2/exercises/run.cmd?exerciseld=mbase://atuf_test/Ordnen_Clemens_th/Ordnen_Clemens&bookId=atuf_test-collinfo

ActiveMath

Übung

Die Schüler sollten 4,

$\frac{12}{24}$, $\frac{16}{2}$ und $\frac{8}{12}$ der Größe nach ordnen. Dabei gingen sie unterschiedlich vor:

Clemens wandelte 4 um und erhielt $\frac{4}{1}$, $\frac{12}{24}$, $\frac{16}{2}$, $\frac{8}{12}$. Dann ordnete er die neuen Brüche nach den Zählern.

Nadine kürzte die Brüche $\frac{12}{24}$, $\frac{16}{2}$, $\frac{8}{12}$ und erhielt 4, $\frac{1}{2}$, $\frac{8}{1}$, $\frac{2}{3}$. Dann ordnete sie die neuen Brüche nach den Nennern und stellte die 4 als größte Zahl ans Ende.

Anne wandelte 4 um und erweiterte die Brüche $\frac{12}{24}$ und $\frac{8}{12}$, kürzte und erweiterte $\frac{16}{2}$ und erhielt $\frac{24}{6}$, $\frac{24}{48}$, $\frac{24}{3}$, $\frac{24}{36}$. Dann ordnete sie die neuen Brüche nach den Nennern.

Jannek wandelte 4 um und erhielt so $\frac{4}{1}$, $\frac{12}{24}$, $\frac{16}{2}$, $\frac{8}{12}$. Dann ordnete er die neuen Brüche nach den Nennern.

Klicke an, wer es **richtig** gemacht hat.

Weiter

Figure 5.2: Typical exercise in the Atuf content delivered by the ActiveMath system

5.2 Difficulty Estimation

After filtering out exercises with lower event count than the number of ability clusters that were used ($k = 10$), 76 exercises from the training set events of the HTW data and 24 exercises from the training set events of the Atuf data were calibrated, refer to Figure 5.3 for an example of the exercise calibration via sigmoid fitting. The summaries from the exercise calibration process are listed in Table 5.3 and Table 5.4. The number of supporting events for an event is the count of the previous events corresponding to the same student and concept as the current event.

Table 5.3 Item calibration process – summary (HTW data)

	Median	Mean
Event Count per Exercise	82	112.91
Nr of Supporting Events per Event	4.84	5.49

Table 5.4 Item calibration process – summary (AtuF data)

	Median	Mean
Event Count per Exercise	108	93.75
Nr of Supporting Events per Event	3.26	3.3

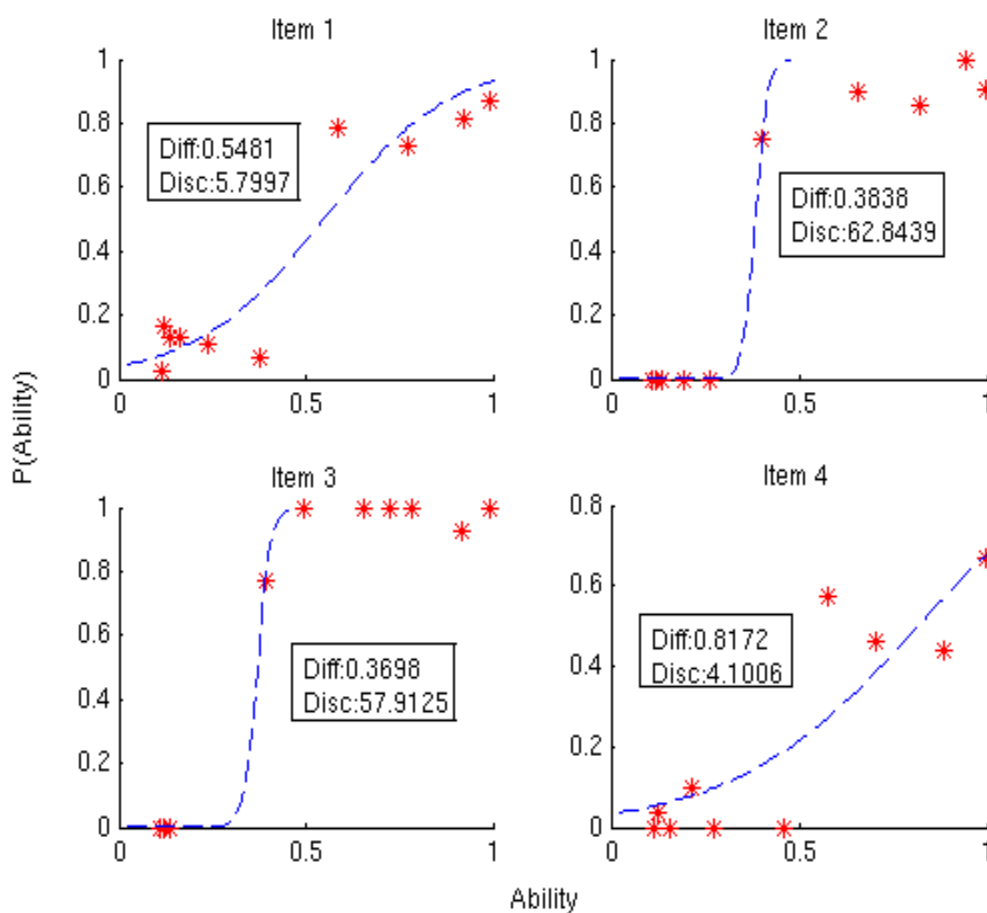


Figure 5.3: Item calibration – sigmoid fitting examples

In Figure 5.3, the 10 red points are the observations in the form of student ability and probability of successful solution pairs and the blue dotted line is the optimal sigmoid curve obtained to fit the observations.

Table 5.5 describes the summary of the difficulty metadata estimates. Figures 5.4 and Figure 5.5 show the histogram visualization of the exercise difficulty metadata for the HTW and AtuF datasets respectively.

Table 5.5 Exercise difficulty metadata – summary

Dataset	Median	Mean	Standard Deviation
HTW	0.5202	0.6189	0.2691
AtuF	0.4470	0.5532	0.2114

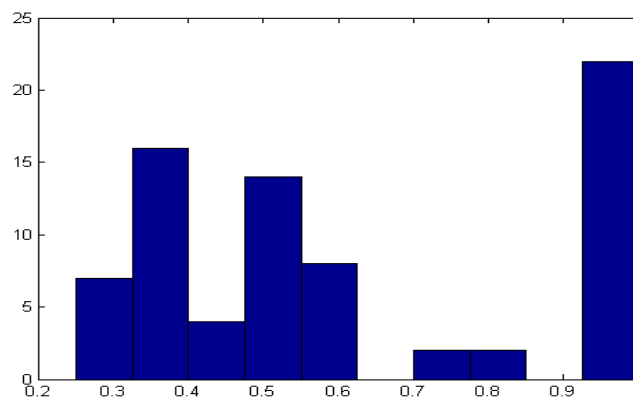


Figure 5.4: Histogram representation of difficulty metadata (HTW)

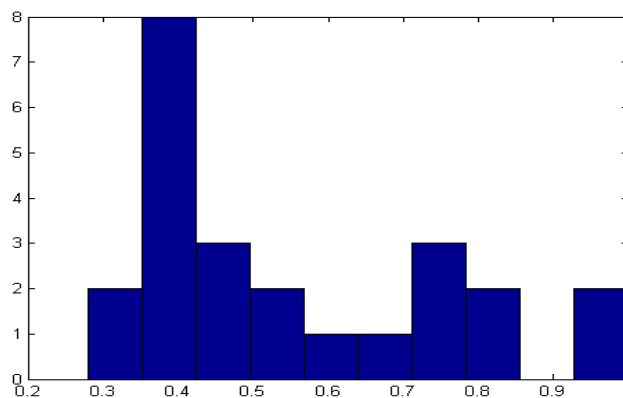


Figure 5.5: Histogram representation of difficulty metadata (AtuF)

5.3 Testing Calibrated Difficulty

The predictive accuracy of our approach is measured using the confusion matrix as described in Table 5.7. The predictive values, predictive rates and F-1 measures for the positive cases, negative cases and the general case using a weighed averaging of the positive and negative cases are computed as described in Table 5.8 & Table 5.9.

Table 5.7 Confusion matrix for predictions

		Actual Result	
		True (P)	False (N)
Prediction	True (P')	True Positives (TP)	False Positives (FP)
	False (N')	False Negatives (FN)	True Negatives (TN)

Table 5.8 Predictive values, predictive rates & F-1 measures (positive & negative)

Positive Predictive Value (PPV)	Positive Predictive Rate (PPR)	Negative Predictive Value (NPV)	Negative Predictive Rate (NPR)
$\frac{TP + 0.5}{TP + FP + 0.5}$	$\frac{TP + 0.5}{TP + FN + 0.5}$	$\frac{TN + 0.5}{TN + FN + 0.5}$	$\frac{TN + 0.5}{TN + FP + 0.5}$
Positive F-1 measure		Negative F-1 measure	
$2 * \frac{PPV * PPR}{PPV + PPR}$		$2 * \frac{NPV * NPR}{NPV + NPR}$	

Table 5.9 Predictive value, predictive rate & F-1 measure (joint – using weighed averaging)

Joint Predictive Value (Weighted Avg.) (JPV)	Joint Predictive Rate (Weighted Avg.) (JPR)
$PPV * \frac{TP + FP}{TP + FP + FN + TN} +$ $NPV * \frac{TN + FN}{TP + FP + FN + TN}$	$PPR * \frac{TP + FN}{TP + FP + FN + TN} +$ $NPR * \frac{TN + FP}{TP + FP + FN + TN}$
Joint F-1 measure	
$2 * \frac{JPV * JPR}{JPV + JPR}$	

Using the same student modeling component (KT model), a comparison between the predictions of the proposed approach (KT + IRT), the predictions obtained for manually annotated exercise difficulties (KT + Annotated exercise difficulty) and the predictions obtained using only student modeling without any exercise difficulty measure (Only KT) was conducted. For the cases where the exercise difficulty was used, a correct solution was predicted if the probability of the student having mastered the underlying concept of the exercise is greater than or equal to the difficulty level of the exercise (The exercise difficulty is measured in a probabilistic scale for both the IRT based estimation and the manual annotation). For the case where no exercise difficulty was used, a correct solution was predicted if the probability of the student having mastered the underlying concept of the exercise is greater than or equal to 0.5. The predictions were restricted to only correct and wrong solutions and don't consider partially correct solutions.

5.3.1 HTW test set predictions

A total of 74 of the 76 calibrated items had events in the HTW test set data, the prediction results for which are listed in the Table 5.10. Based on the predictions results the predictive values, predictive rates and F-1 measures were calculated and are listed in Table 5.11 and Table 5.12.

Table 5.10 Confusion matrix for predictions (HTW data)

	True Positives	False Positives	False Negatives	True Negatives
KT + IRT (Thesis approach)	906	159	15	996
KT + Annotated exercise difficulty	911	591	10	564
Only KT (No exercise difficulty)	818	214	103	941

Table 5.11 Predictive values, predictive rates & F-1 measures (positive & negative - HTW data)

	Positive Predictive Value	Positive Predictive Rate	Negative Predictive Value	Negative Predictive Rate
KT + IRT (Thesis approach)	0.8508	0.9837	0.9852	0.8624
KT + Annotated exercise difficulty	0.6067	0.9891	0.9826	0.4885
Only KT (No exercise difficulty)	0.7927	0.8882	0.9014	0.8148
	Positive F-1 measure		Negative F-1 measure	
KT + IRT (Thesis approach)	0.9124		0.9197	
KT + Annotated exercise difficulty	0.7521		0.6526	
Only KT (No exercise difficulty)	0.8377		0.8559	

Table 5.12 Predictive value, predictive rate & F-1 measure (joint – using weighed averaging – HTW data)

	Joint Predictive Value (Weighted Avg.)	Joint Predictive Rate (Weighted Avg.)	Joint F-1 (Weighted Avg.)
KT + IRT (Thesis approach)	0.9328	0.9290	0.9308
KT + Annotated exercise difficulty	0.7628	0.7573	0.7599
Only KT (No exercise difficulty)	0.8763	0.8672	0.8717

5.3.2 AtuF test set predictions

A total of 20 of the 24 calibrated items had events in the AtuF test set data, the prediction results for which are listed in the Table 5.13. Based on the predictions results the predictive values, predictive rates and F-1 measures are calculated and are listed in Table 5.14 and Table 5.15.

Table 5.13 Confusion matrix for predictions (AtuF data)

	True Positives	False Positives	False Negatives	True Negatives
KT + IRT (Thesis approach)	347	21	6	127
KT + Annotated exercise difficulty	353	30	0	118
Only KT (No exercise difficulty)	337	24	16	124

Table 5.14 Predictive values, predictive rates & F-1 measures (positive & negative - AtuF data)

	Positive Predictive Value	Positive Predictive Rate	Negative Predictive Value	Negative Predictive Rate
KT + IRT (Thesis approach)	0.9430	0.9830	0.9551	0.8586
KT + Annotated exercise difficulty	0.9218	1	1	0.7980
Only KT (No exercise difficulty)	0.9336	0.9547	0.8861	0.8384
	Positive F-1 measure		Negative F-1 measure	
KT + IRT (Thesis approach)	0.9626		0.9043	
KT + Annotated exercise difficulty	0.9593		0.8877	
Only KT (No exercise difficulty)	0.9440		0.8616	

Table 5.15 Predictive value, predictive rate & F-1 measure (joint – using weighed averaging – AtuF data)

	Joint Predictive Value (Weighted Avg.)	Joint Predictive Rate (Weighted Avg.)	Joint F-1 (Weighted Avg.)
KT + IRT (Thesis approach)	0.9379	0.9389	0.9384
KT + Annotated exercise difficulty	0.9285	0.9310	0.9297
Only KT (No exercise difficulty)	0.9016	0.8984	0.8999

The results from this evaluation show a gain in the accuracy of student performance predictions with the use of exercise difficulty calibrated by the proposed approach as compared to predictions based on manually annotated exercise difficulties or predictions considering all exercises to have the same difficulty. The higher predictive value and lower predictive rate on the negative predictions as compared to the positive predictions when using exercise difficulty suggest a general bias in the exercise difficulty measure towards the lower end of the scale. This bias is observed to be stronger for manually annotated exercise difficulties as compared to the exercise difficulties calibrated by the proposed approach, which translates to manual annotation of exercises difficulty showing to be easier than they actually are. These observations on the exercise difficulty are further supported by the more balanced predictions when we consider only student modeling without the use of exercise difficulty.

5.4 Discrimination Estimation

Table 5.6 describes the summary of the discrimination metadata estimates. Figures 5.6 & Figure 5.7 show the histogram visualization of the exercise discrimination metadata for the HTW and AtuF datasets respectively.

Table 5.6 Exercise discrimination metadata - summary

Dataset	Median	Mean	Standard Deviation
HTW	67.1975	99.8641	87.1990
AtuF	50.1499	49.8432	35.2305

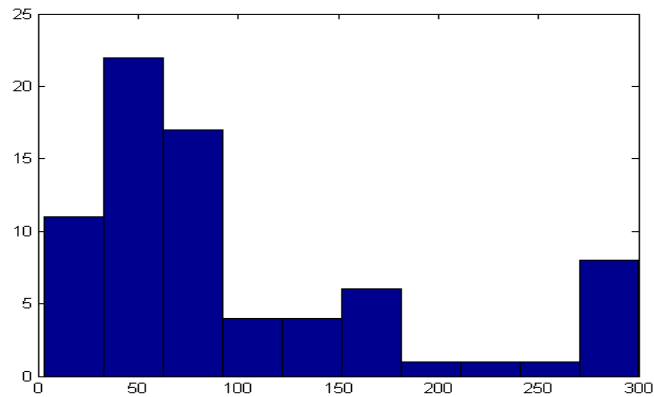


Figure 5.6: Histogram representation of discrimination metadata (HTW)

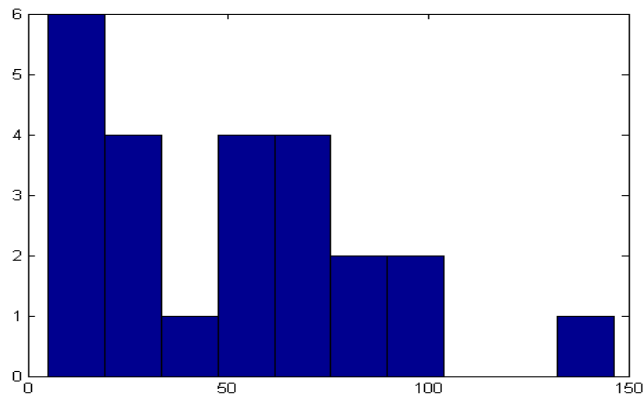


Figure 5.7: Histogram representation of discrimination metadata (AtuF)

The estimates of the discrimination metadata was bounded between 0 and 300, since a negative discrimination would mean an exercise that is solved successfully by students without the mastery of the addressed concept but incorrectly by students who have mastery of the concept, which is very unnatural and since it was important to have an upper bound to avoid having the discrimination value iteratively approach infinity, it was bounded by 300 which is a fairly high discrimination value.

It was observed that the higher discrimination estimates were often accompanied by high exercise difficulty estimates, which was because all the students that had answered the exercise got it wrong. Since the sigmoid curve we fit to this data has to achieve the highest probability for the highest point on the ability scale, it is forced to an abrupt increase in its value at the highest ability level, hence resulting in a very high discrimination estimate.

Chapter 6

Conclusions and Discussion

Having observed a gain in the predictive accuracy with the exercise difficulty estimation using the proposed approach over the student performance predictions using manually annotated exercise difficulty and the predictions using just the student model without any exercise difficulty information, in this section we proceed to further discuss the evaluation framework, the data used, some comparisons of the proposed approach and possible future directions that could be explored as a continuation of the work presented in the thesis.

6.1 Evaluation Design

The binary classification scheme used in the evaluation of the exercise difficulties estimated by the proposed approach is intuitive and leads to a straight forward interpretation when the true difficulties of the exercises are unknown, which was the case with the evaluation data used. This evaluation scheme could further be improved by taking into account the gap between the student ability and exercise difficulty, especially for the false predictions, when calculating a score for the prediction accuracy.

Other important aspects involving the evaluation of the exercise difficulties are:

- All three techniques (the proposed approach, the manually annotated difficulty estimates & the no difficulty estimation) that are compared in the evaluation used the same student modeling scheme (KT model) with the exact same parameters for a fair evaluation.
- The manually annotated exercise difficulties based on a discrete scale is compared against the continuous scale of the estimations obtained from the proposed approach. Hence the accuracy of the manual annotations is at a disadvantage compared to the estimations obtained by the proposed approach.

6.2 Differences in Evaluation Data

The overall higher predictive accuracies for the Atuf data as compared to HTW data could be due to the higher number of exercises per concept (Atuf (20 exercises for 5 concepts) vs. HTW (74 exercises for 33 concepts)). This observation is supported by the greater difference in the HTW exercises as compared to the Atuf exercises when considering the accuracy gap between the predictions using the student modeling alone and the predictions using the proposed approach. Hence, we note that the concept count to exercises count ratio of the dataset could serve to be an important consideration in accounting for the differences in the accuracy of student performance prediction between different datasets, especially when a Bayesian knowledge tracing scheme is employed for the student modeling.

The lower difference in the joint F1 measures between the HTW and the Atuf data using the proposed approach as compared to the same using only the student modeling indicated that the proposed approach has a greater gain in predictive accuracy on the HTW data as compared to the Atuf data. The reason for this could be the higher average number of supporting events for each event in the HTW data, which would justify the logarithmic weighing scheme used to indicate the importance of an event in estimating its exercise difficulty based on the support it has from past events corresponding to the same student and concept. It may also be interesting to note that mean event count for an exercise in the HTW data was greater but the median of the event count for an exercise in Atuf data was greater.

It is readily observed that the manually annotated difficulties for the Atuf exercises are much more accurate than those for the HTW exercises. The reason for this could be that the Atuf project was a controlled study performed by a group of educational psychologists who took great care in annotating the test exercise difficulties as compared to the HTW exercises which were meant for training and probably were annotated by a domain expert without much experience in educational psychology.

The overall higher predictive accuracies for the Atuf data could also be due to the fact that it was a controlled study, reducing the possible noise in the observed data. The students in the Atuf study had to solve all the test exercises in a short time frame as compared to the students in the HTW data who were free to train on the exercises without such time session restrictions, hence giving them the opportunity to learn the exercise related concepts through external means.

6.3 Approach Comparison

The approach in this thesis could be compared to Performance Factor Analysis (PFA) (Pavlik et al., 2009), which reconfigures Learning Factor Analysis (LFA) based cognitive model (Cen et al., 2006), using the Rash model which the IRT is based on. Evaluation of PFA as compared to KT based modeling have shown improvements in accuracy, especially when multi knowledge component exercises are considered (Cen et al., 2006). Though another study by Gong et al. (2010) comparing PFA with KT based approaches that did not use exercise difficulty in prediction failed to show much difference in accuracy improvement with PFA, the authors hypothesize that PFA works better in circumstances where exercises for a particular knowledge component vary greatly in difficulty.

This thesis would be incomplete without addressing the traditional IRT based calibration techniques originally used in Computer Adaptive Testing (CAT). A search for IRT based exercise calibration techniques used outside of the field of EDM reveals the following approaches towards joint estimation of item and ability IRT parameters.

1. Joint maximum likelihood (JML), exemplified by LOGIST (Wingersky, 1983; Wingersky, Barton, & Lord, 1982)
2. Marginal maximum likelihood (MML), exemplified by BILOG (Mislevy & Bock, 1983)
3. Bayesian estimation (Swaminathan & Gifford, 1982; 1985; 1986)
4. Markov chain Monte Carlo approaches (MCMC)

The maximum likelihood estimation (MLE) based approaches guarantee only local minima, due to the use of gradient decent based numerical methods to solve a non-convex problem, which is usually the case in educational data due to noise created by erratic or negatively discriminating items and student actions that are inconsistent with the student's latent trait, like guessing the answer to an item. Any improvements the MML has over JML depend greatly on the chosen prior ability distribution, since MML just separates the ability parameter optimization from the item parameter optimization using the Expectation Maximization (EM) algorithm (Mislevy & Stocking, 1989). The Bayesian hierarchical model based estimation approaches, being more robust in estimating the prior information from the observed data have empirically outperform MLE approaches, especially on small datasets (Lord, 1986; Swaminathan & Gifford, 1982; 1986). MCMC methods have typically been used in more complex IRT based modeling, where estimation of item and ability at the same time is desired, but are known to have a slower execution time in spite of employing randomization techniques in the estimation of the joint posterior distribution function of all the parameters (Patz & Junker, 1999).

In our approach, we use data mining to capitalize on the prior information we have about the items, specifically the item clustering based on the knowledge component they address. This enables a more robust estimation of the student ability parameter. The selection of the 4 parameters in our KT based student modeling plays a vital role in the predictive performance of our approach because of the strong reliance on the accuracy of the student modeling. By intuitively restricting the values of the KT parameters, the optimal KT parameters can be estimated based on the observed test data using a histogram based search. Relying on the robustness of our student ability estimates we resort to a simple averaging based computation of the conditional probability of a correct solution to an item, given the ability level of the student. The most likely item parameters are then estimated based on this modeled posterior distribution. Another advantage we gain by using data mining is a detailed account of the history of the data supporting our estimations.

Our approach could be compared to the Marginal Maximum Likelihood (MML) technique, in the sense that we choose a prior distribution on the student parameters in the form of the KT parameters and then solve for the most likely item parameters, given the observations. The estimates can be considered as being finer grained though, because we only marginalize along the items sharing the same knowledge component rather than along all the items. And the curve fitting using a weighted average based least square distance minimization is better equipped in mitigating the influence of outliers and the local minima problem.

6.4 Future Work

It is planned to implement the proposed approach as a tool to calibrate the exercise difficulties for the content delivered by the ActiveMath (ActiveMath, n.d.) e-learning system. A possible extension we could consider is designing a calibration quality measure for the exercises calibrated using our approach which takes into account information on the ability range of the supporting events we register for each exercise, along with the events count and the average number of supporting events per event.

The ActiveMath metadata scheme contains a field that describes the exercise purpose, by specifying if an exercise is suitable for training, testing or both. The discrimination metadata obtained using the proposed approach in the thesis can be helpful for distinguishing between a training exercise and a test exercise when the information is unavailable, since test exercises typically have a higher discrimination value than training exercises.

The discrimination metadata which provides valuable information on the strength of an exercise in differentiating between the students who have mastered the addressed concept and the students who have not, can be used to further improve on the student performance predictions. This can be done by developing a measure based on the discrimination metadata that could be compared against the difference between the student ability and the exercise difficulty. This would in simple terms, compare the distance of a students' current ability level in

a concept from the mean ability level required to solve the exercise with the variance of the ability levels of students who have solved the exercise successfully in the past.

It would be interesting to use the information on the discriminative characteristic of exercise, that our approach provides us with in analyzing the student behavior on the different groups of exercises. This information could possibly help in inferring the group of exercises that best facilitate learning along with other easily available information such as time taken to answer the exercises.

Another interesting area to explore could be the analysis of the metadata of the exercises sharing the same knowledge component to have content specific inferences on the knowledge component.

An extension of our approach to multi step and multi concept exercises could include considering a sigmoid mixture model and evaluating its performance against Markov Chain Monte Carlo (MCMC) based calibration techniques, which have traditionally been used to model complex exercises with multiple concepts.

References

ActiveMath (n.d.). <http://www.activemath.org/>

Adaptive tutorial Feedback (AtuF) project (n.d.). <https://sites.google.com/site/atufproject/project-updates/to-dos/project-definition>

Amershi, S., Conati, C. (2009). Combining Unsupervised and Supervised Classification to Build User Models for Exploratory Learning Environments. *Journal of Educational Data Mining* 1(1), 18–71.

Baker, R.S.J.d, & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, Vol. 1, Issue 1, pp. 3-17.

Baker, R.S.J.d. (2007) Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems. *Proceedings of ACM CHI 2007: Computer-Human Interaction*, 1059-1068.

Baker, R.S.J.d. (2010). Data Mining. In Penelope Peterson, Eva Baker, & Barry McGaw (Eds.), *International Encyclopedia of Education* (3rd edition). Oxford, UK: Elsevier.

Baker, R.S.J.d. (2010). Mining Data for Student Models. In Nkambou, R. et al. (Eds.). (2010). *Advances in Intelligent Tutoring Systems*, SCI 308, pp. 323–337. Berlin Heidelberg: Springer-Verlag.

Baker, R.S.J.d., Corbett, A.T., Roll, I., & Koedinger, K.R. (2008). Developing a Generalizable Detector of When Students Game the System. *User Modeling and User-Adapted Interaction*, Vol. 18(3), pp. 287–314.

Barnes, T. (2005). The q-matrix method: Mining student response data for knowledge. In *Proceedings of the AAAI-2005 Workshop on Educational Data Mining*.

Beck, J., Stern, M., & Haugsjaa, E. (1996). Applications of AI in Education. *Magazine Crossroads - Special issue on artificial intelligence*, Vol. 3, Nr. 1.

Beck, J.E. & Mostow, J. (2008). How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, 353-362.

Beck, J.E., & Sison, J. (2006). Using Knowledge Tracing in a Noisy Environment to Measure Student Reading Proficiencies. *International Journal of Artificial Intelligence in Education*, Vol. 16, No. 2, pp. 129-143.

Bloom B.S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, Vol. 13, No. 6, pp. 4-16.

Cen, H., Koedinger, K., & Junker, B. (2006). Learning factors analysis a general method for cognitive model evaluation and improvement. In *Intelligent Tutoring Systems*. Vol. 4053. Springer Berlin Heidelberg, 164–175.

Cocea, M., HersHKovitz, A. & Baker, R.S.J.D. (2009). The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate? In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 507-514.

Coleman, T.F. and Li, Y. (1996). An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds. *SIAM Journal on Optimization*, Vol. 6, pp. 418–445.

Conejo, R.; Guzmán, E.; Millán, E.; Pérez-de-la-Cruz, J. L., Trella, M., & Ríos, A (2004). SIETTE: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education*, 14, 29-61.

Corbett, A.T., & Anderson, J.R. (1995). Knowledge Tracing Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4, 253-278.

Desmarais, M.C. and Pu, X. (2005). A Bayesian Student Model without Hidden Nodes and Its Comparison with Item Response Theory. *International Journal of Artificial Intelligence in Education* 15, 291-323.

ETS. Educational Testing Service (n.d.). <http://www.ets.org>

Feng, M., Heffernan, N., Beck, J.E., & Koedinger, K. (2008). Can we Predict which Groups of Questions Students will Learn from? In *Proceedings of the First International Conference on Educational Data Mining*, 218-225.

Gong, Y. & Beck, J. E. (2011). Items, skills, and transfer models: which really matters for student modeling? In *Proceedings of the Fourth International Conference on Educational Data Mining*, 81-90.

Gong, Y., Beck, J.E., & Heffernan, N.T. (2010). Comparing Knowledge Tracing and Performance Factor Analysis by Using Multiple Model Fitting Procedures. Alevén, V., Kay, J., and Mostow, J. (Eds.), *ITS 2010, Part I, LNCS 6094*, pp. 35–44. Berlin Heidelberg: Springer-Verlag.

Guzmán, E., & Conejo, R. (2005). Towards Efficient Item Calibration in Adaptive Testing. In *Proceedings of the 10th international conference on User Modeling*, pp. 402-406.

HTW. Hochschule für Technik und Wirtschaft, Saarbrücken (n.d.). http://www.htw-saarland.de/index_html/

Jeong, H., & Biswas, G. (2008). Mining Student Behavior Models in Learning-by- Teaching Environments. In Proceedings of the 1st International Conference on Educational Data Mining, 127-136.

Koedinger, K., Baker, R., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A Data Repository for the EDM Community: The PSLC Datashop. Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R.S.J.D. (Eds.). Handbook of Educational Data Mining. Boca Raton, FL: CRC Press.

Langley, P., & Ohlsson, S. (1984) Automated Cognitive Modeling. Proceedings of the National Conference on Artificial Intelligence, pp.193-197.

Li, N., Matsuda, N., Cohen, W.W., & Koedinger, K.R. (2011). A Machine Learning Approach for Automatic Student Model Discovery. In Proceedings of the Fourth International Conference on Educational Data Mining, pp. 31-40.

Lord F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Lord F.M. (1986). Maximum Likelihood and Bayesian Parameter Estimation in Item Response Theory. Journal of Educational Measurement, Vol. 23, No. 2, pp. 157-162.

MacQueen, J. B. (1967). "Some Methods for classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297.

Martin, J., & Van Lehn, K. (1995). Student Assessment Using Bayesian Nets. International Journal of Human Computer Studies, Vol. 42, pp. 575-591.

Math-Bridge Education Solution (n.d.). <http://www.activemath.org/>

Mislevy, R. J., & Bock, R. D. (1983). BILOG: Item Analysis and Test Scoring with Binary Logistic Models [Computer program]. Mooresville IN: Scientific Software, Inc.

Mislevy, R.J., & Stocking M.L. (1989). A Consumer's Guide to LOGIST and BILOG. Applied Psychological Measurements, Vol. 13, No. 1, pp. 57-75.

Nkambou, R., Bourdeau, J., & Mizoguchi, R. (Eds.) (2010). Advances in Intelligent Tutoring Systems. Berlin Heidelberg: Springer-Verlag.

Nkambou, R., Nguifo, E.M., Couturier, O., Fourier-Vigier, P. (2007). Problem-Solving Knowledge Mining from Users' Actions in an Intelligent Tutoring System. In: Kobti, Z., Wu, D. (eds.) Canadian AI 2007. LNCS (LNAI), vol. 4509, pp. 393–404. Springer, Heidelberg.

Nwana, H.S. (1990). Intelligent Tutoring Systems: an overview. Artificial Intelligence Review, Vol. 4, Nr. 4, pp. 251-277.

- Pardos, Z.A., & Heffernan, N.T. (2009). Determining the Significance of Item Order in Randomized Problem Sets. In Proceedings of the Second International Conference on Educational Data Mining, pp.11-121.
- Patz, R.J.; Junker, B.W. (1999). A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models. *Journal of Education and Behavioral Statistics*, Vol. 24(2), pp. 146-178.
- Pavlik, P.I., Cen, H., & Koedinger, K (2009). Performance Factors Analysis - A New Alternative to Knowledge. In Proceedings the 14th International Conference on Artificial Intelligence in Education, pp. 531–538.
- Rai, D., Gong, Y., & Beck, J.E. (2009). Using Dirichlet priors to improve model parameter plausibility. In Proc. Int. Conf. Educ. Data Mining, pp. 141–150.
- Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K. R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T.E., Upalekar, R., Walonoski, J.A., Macasek.M.A. & Rasmussen, K.P. (2005). The Assistment project: Blending assessment and assisting. In C.K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.) Proceedings of the 12th Artificial Intelligence in Education, Amsterdam: ISO Press. pp. 555-562.
- Romero, C., & Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. *Expert Systems with Applications* 33, 125-146.
- Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of Art. In *IEEE Transactions on Systems, Man and Cybernetics*, Vol.40, No. 6, pp. 601-618.
- Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R.S.J.D. (Eds.). (2011). *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press.
- Rus, V., Lintean, M., & Azevedo, R. (2009). Automatic detection of student mental models during prior knowledge activation in MetaTutor. In Proc. Int. Conf. Educ. Data Mining, Cordoba, Spain, pp. 161–170.
- Shute, V., & Towle, B., (2003). Adaptive E-Learning. *Educational psychologist*, Vol. 38(2), pp. 105-114.
- Shute, V.J. (1995). SMART: Student modeling approach for responsive tutoring. *User Modeling and User-Adapted Interaction*, 5(1), 1-44.
- Sleeman, D.H., Langley, P., & Mitchell, T.M. (1982) Learning from solution paths: an approach to the credit assignment problem. *AI Magazine*, 3(1), 48-52.
- Swaminathan, H., & Gifford, J. A. (1982). Bayesian Estimation in the Rasch Model. *Journal of Educational and Behavioral Statistics* Statistics, Vol. 7(3), pp.175-192.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian Estimation in the Two-Parameter Logistic Model. *Psychometrika*, Vol. 50, pp. 349-364.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian Estimation in the Three-Parameter Logistic Model. *Psychometrika*, Vol. 51(4), pp. 589-601.

Tang, T., & McCalla, G. (2005). Smart recommendation for an evolving e-learning system. *International Journal on E-Learning*, Vol. 4, Issue 1, pp. 105–129.

Thai-Nghe, N., Horváth, T., & Schmidt-Thieme, L. (2011). Factorization Models for Forecasting Student Performance. In *Proceedings of the Fourth International Conference on Educational Data Mining*, 11-20.

van der Linden, W.J., & Hambleton, R.K. (Eds.) (1997). *Handbook of Modern Item Response Theory*. New York, NY: Springer Verlag.

Walonoski, J.A., Heffernan, N.T. (2006). Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) *ITS 2006*. LNCS, vol. 4053, pp. 382–391. Springer, Heidelberg.

Walonoski, J.A., Heffernan, N.T. (2006). Prevention of Off-Task Gaming Behavior in Intelligent Tutoring Systems. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) *ITS 2006*. LNCS, vol. 4053, pp. 722–724. Springer, Heidelberg.

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2011). Acquiring Item Difficulty Estimates: a Collaborative Effort of Data and Judgment? In *Proceedings of the Fourth International Conference on Educational Data Mining*, 121-128.

Wauters, K., Desmet, P., & van den Noortgate, W. (2010). Adaptive Item-Based Learning Environments Based on the Item Response Theory: Possibilities and Challenges. *Journal of Computer Assisted Learning*, 26(6), 549-562

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton NJ: Educational Testing Service.

Wingersky, M.S. (1983). LOGIST: A Program for Computing Maximum Likelihood Procedures for Logistic Test Models. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver BC: Educational Research Institute of British Columbia.

Witten, I.H., & Frank, E. (1999). *Data mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Fransisco, CA: Morgan Kaufmann.

Zaïane, O.R. (2002). Building a Recommender Agent for e-Learning Systems. In *Proceedings of the International Conference on Computers in Education - 2002*, Vol. 1, pp. 55-59.

Hershkovitz, A., Nachmias, R. (2008). Developing a Log-Based Motivation Measuring Tool. In: Proceedings of the First International Conference on Educational Data Mining, pp. 226–233.