LEARNING BY DOING AND LEARNING WHEN DOING

Dovetailing E-Learning and Decision Support with a Data Mining Tutor

Klaus P. Jantke, Steffen Lange
German Research Center for Artificial Intelligence, Saarbrücken, Germany
Email: {jantke,lange}@dfki.de

Gunter Grieser, Peter Grigoriev
Technical University of Darmstadt, Dept. Informatics, Darmstadt, Germany
Email: {gunter.peter}@informatik.tu-darmstadt.de

Bernhard Thalheim, Bernd Tschiedel
Technical University of Cottbus, Dept. Informatics, Cottbus, Germany
Email: {thalheim,tschiedel}@informatik.tu-cottbus.de

Keywords: E-Learning, Knowledge Discovery, Data Mining, Decision Support, Learning on Demand

Abstract: In this paper, e-learning meets decision support in enterprises’ business practice. This presentation is based on an on-line e-learning system named DaMiT for the domain of knowledge discovery and data mining. The DaMiT system was primarily developed for technology enhanced learning in German academia. It is now on the cusp of entering training on demand in enterprises. Stand-alone e-learning seems quite unrealistic and does not meet the needs of industries. It is very unlikely that employees fully loaded with work take a detour to study theories of whatever sort. More likely, they are willing to engage in studies whenever the need derives directly from their practical work. In those cases, they might even be willing to dive into theories. How to dovetail e-learning and enterprise business applications, such that both sides draw a proper benefit?

1 INTRODUCTION

There is no doubt at all that technology enhanced learning is going to change education on all levels ranging from ordinary schools over universities to professional training and life-long learning. The process is boosted by the Internet in pervading the world.

The recently observable progress in the area named e-learning is enormous and ranges from a flood of content (For illustration, the German Federal Ministry for Education and Research, BMBF, has put about 200 Mio. Euro into 100 joint projects to develop content for academic e-learning. Another 200 Mio. Euro went into schools and professional education, all this within only 3 years,) to technological innovations. We all are on the cusp of the invention of truly adaptive e-learning systems, based on deep learner modelling, expressive XML-based content representation and, last but not least, flexible, attractive and appealing generation and presentation technologies.

There are still a number of open problems also in technology, but the research and development community is very active.

In industries, however, one observes an obvious reluctance. Employees fully loaded with work tend to restrain from getting involved in extra activities. And the management frequently has understandable reservations about introducing another software system and further diversifying the IT infrastructure.

This situation bears abundant evidence for the need of truly integrating e-learning into the business processes and the IT infrastructures of enterprises. There is not much hope for the success of stand-alone e-learning solutions in the wild.

Last but not least, the questions under discussion are relevant to universities and other academic institutions when pondering about business models and marketing potentials of e-learning cruising for their own position in the market.
2 ALMOST A CREDO

Among many other ideas, the authors are very much in favour of the two phrases which gave the title for the present paper: learning by doing and learning when doing. We keep it short:

It is not really easy to find an agreement among scientists about what knowledge might be or how it might be circumscribed. Dictionaries find explanations like the following: “the facts, information, understanding and skills that a person has acquired through experience and education”.

When learning is about bringing knowledge to the learner and when e-learning is about doing this by means of computer networks and dedicated software systems, how to send information, understanding or even skills through the Internet? Obviously, there is no way to encode all this at the one side into bits and bytes, pump it through the network and decode it at the other side.

Knowledge has to be acquired by the learner, and this may be sometimes a laborious process which requires substantial user activity. In quite interesting cases, different learners at computer terminals dealing with e-learning tasks may acquire different knowledge from the same bits and bytes arriving at their respective terminals. Clearly, the knowledge is not sitting in the bits and bytes, but it is (re)constructed by the learner in a process of human-machine interaction, possibly, including also off-line work to some extent – learning by doing.

Doubtlessly, learning non-trivial things may be time-consuming and sometimes straining. Therefore, it does not easily fit into a tight work schedule. In particular, theoretical foundations often seem superfluous, if the learner’s main interest is in solving certain application problems.

In industrial settings, it fits best when learning activities are initiated in response to practical needs. If a problem has to be solved and several approaches are available, it may be handy to have information available to learn about the strength and weaknesses of these approaches – learning when doing.

So, what is the present paper about? Outgoing from the authors’ work in e-learning and based on the authors’ and their colleagues’ e-learning system DaMiT, the importance of learning by doing is discussed in some detail and exemplified. The doing-oriented approach of DaMiT is taken as a basis for introducing “learning on demand” into enterprises where management decisions have to be based on comprehensive and usually distributed data and where understanding complex data is essential. The authors’ proposal is to place (links to) the tutoring system DaMiT in an enterprise’s IT infrastructure such that learning when doing is enabled.

Future work should lead to integrated systems where a system like DaMiT does not appear as an outsider. Enterprise application integration abbreviated EAI (Linthicum, 2000) is, perhaps, the first step to integrate e-learning systems like DaMiT.

To the authors’ very best knowledge, there has not been any similar approach before, at least about the content of knowledge discovery and data mining, except their own first steps reported from a rather different perspective in (Jantke et al., 2003).

3 THE DaMiT BACKGROUND

The DaMiT system has been developed within a joint project carried out by groups in 10 German universities in collaboration with the DFKI which is the German Research Center for Artificial Intelligence.

Originally, the DaMiT system was set up as an academic system for e-learning in institutions of higher education in Germany. Therefore, the first system version (cf. the link Lernsystem under the address http://DaMiT.dfki.de) and the vast majority of early publications like (Degel, 2003), (Grieser et al., 2003), (Grieser and Grigoriev, 2003), (Memmel, 2003) and others are in German, only.

Though the DaMiT systems do contain a rather balanced collection of theoretical background and application oriented knowledge about algorithms, it has a clear focus on motivating users by attractive application cases. On the top level of the DaMiT system, registered students as well as anonymous visitors just dropping in can find case studies of a realistic size and of practical relevance.

Figure 1: Case Studies in the DaMiT Tutor System

Four of these case studies are in fact the detailed descriptions of award winning solutions of an international competition in knowledge discovery and data mining, the so-called Data Mining Cup (cf. http://www.data-mining-cup.de ).
Those award winning problem solutions in application areas like e-mail spam filtering and costumer loyalty forecasting for direct marketing, for instance, are truly relevant and reflect scientifically and technically the best practice world-wide.

Case studies may be used as a problem-oriented entry into the e-learning system. Other entry points are found via the glossary or, traditionally, through the table of contents.

The authors refrain from a complete introduction which may be worth some separate paper, but prefer a sketch focussing peculiarities of the domain. The following questions might be helpful as a guideline.

What is data mining about? What makes studies in data mining different? How to abandon the ivory tower of academic studies?

Some authors and especially those advertising data mining software prefer the perspective at data mining as digging for golden nuggets in complex and usually distributed data bases of enterprises, governmental institutions or research institutes. The treasures promised are insights sitting in the data, but hiding there and being unknown to the user. There are many reasons for adopting this view, and one reason might be in marketing. You offer a tool to the users and tell them just to dig for the treasures to become knowledgeable and, thus, successful and even rich.

In the authors opinion, this perspective is only suitable in a restricted number of rather simple situations.

In contrast to the above-mentioned oversimplifying approach sketched, data mining truly deals with building models – mathematical models, to say it very explicitly – over complex data. Those models may be classifiers or clustering algorithms, e.g., and those classifiers, for instance, may be decision trees or rule sets, e.g. To go into more sophisticated detail, the tests employed by decision trees might be tests for text pattern structures, in application domains like biochemistry, e.g. (Arikawa et a., 1993).

Models to be generated over a possibly longer cycle of process steps are expected to have certain predictive power. They are build over some given data and to be exploited in the future over data that are unknown so far.

Thus, data mining is a creative process, to some extent. Humans and machines co-operate in a way that both sides are bringing in their mutual strength. Computers search paramount model spaces, generate complex models, evaluate them over given data and visualize intermediate results. In turn, humans restrict search spaces by means of heuristics, intuitively set preferences and guide the overall process. Data mining is both a science and an art.

The DaMiT system is designed to study the science of data mining and to experience the art.

Data mining studies are complicated, because one inevitably needs to dovetail theoretical investigations and experimental work.

Students will never read a data mining text book from the beginning to the end, and this is why the time of text book alone is over. You do need some support of practicing data mining. An e-learning system provides the appropriate environment.

The crux is that different individuals do prefer respectively need a different way of changing learning modes. The tastes and styles in learning differ enormously. Appropriate educational design needs an anticipation of a variety of learning scenarios. In DaMiT, this is implemented to extensive story boarding (cf. (Jantke et al., 2003).

As data mining by its very nature is not only of interest to students, but relevant to industries, governmental institutions and science, the DaMiT system has been opened to world-wide access.

At present, some key content which is open to all registered students can be accessed by anonymous users through e-payment only. This is the current way to avoid frequent contracting. In the future, there may be other ways of commercial access.
4 ACTIVE LEARNING IN DaMiT

The present chapter is devoted to a detailed discussion of learning by doing in the DaMiT system. The following two chapters are showing how to exploit the doing-oriented features of DaMiT for learning when doing in industrial settings, in governmental working environments or in research institutes, respectively. To bridge this gap is the aim of the present submission.

4.1 Observational Learning

Recall that data mining may be seen both as a science and an art. When practicing the art of data mining, a quite substantial amount of the underlying knowledge is implicit. But how to transmit implicit knowledge by means of e-learning? This is a particularly tough question if the teachers are not always aware of the knowledge they are giving away when being engaged in teaching. Sometimes, one says you just need to get some feeling about it. This is very important in areas like data mining.

The problem of implicit knowledge is even more important when the domain is a rather young one and results are not matured, established publications do not yet exist and teachers are not experienced in how to deal with the material. DaMiT contains some of such highly up-to-date stuff.

The DaMiT system is equipped with several “playgrounds” where learners can experience those phenomena which are rather difficult to deal with explicitly. Figure 3 above shows an applet which allows for exploring the way in which text patterns according to (Angluin, 1980) generate strings resp. sets of strings. Dealing with patterns and pattern languages is not custom in the data mining community, though patterns play a crucial role in important decision tree applications (Katoh et al., 2003).

It is surely one of the highlights in education when learners are able to pose interesting problems to their teachers. Figure 4 displays an applet where a learner can generate decision problems to be solved by the DaMiT system in generating decision trees over regular patterns. The learners provide the input data, i.e. positive and negative examples, and the system generates a certain decision tree with regular patterns serving as tests in the nodes of the tree. The learners can inspect the generated tree and, then, can modify the posed learning problem according to their ideas of how to make the learning task more easy or more difficult to the system. Note that already the tiny problem posed in Figure 4 drives the system to generate a quite interesting decision tree.

Learners experience different phenomena like, e.g., how very small changes in the data result in enormous changes of the decision tree induced or, alternatively, when substantial changes to the data do not change at all the hypothesized classifier.

Observational learning of this type can not easily be substituted by other learning forms.

Data mining always contains a phase of exploration (cf. Figure 2), when users analyze the given data and try to exhibit how to get a clue for attacking a given data mining problem.

QuDA is a professional system for knowledge discovery and data mining which can be accessed trough DaMiT (Grigoriev and Yevtushenko, 2003). In the e-mail spam filtering case study (for details, see the next section) exploring the given data with QuDA provides insights into the relevance of attributes for e-mail classification.

Figure 4: An Applet for Posing Tasks to the System

Figure 5: Concept Analysis with QuDA in DaMiT

The user can activate an attribute name and visualize the relevance of all the activated attributes for the classification task currently under consideration. In the case displayed, the attribute meaning that the phrase “all natural” appears in an e-mail text is related to the classification target in a Hasse diagram. One can add attributes and study their impact visually towards an understanding of how to continue.
4.2 Experiencing True Data Mining

There is not much hope for learning to swim or to ride a bicycle when sitting on a sofa, only. Quite analogously, there is not much hope for learning data mining by reading textbooks or texts of some web-based e-learning system like DaMiT, e.g., only. You need to do it, and you need to do it properly.

The DaMiT system does contain not only case studies as mentioned above, but it also contains what we call competitive exercises. Those exercises are of the quality of practical data mining problems. There is a continuous competition among all learners in finding better and better solutions to these problems.

One can never be sure that a solution found so far will remain best for ever (Strutz and Degel, 2002).

Figure 6: A Competitive Exercise in DaMiT

For tackling a particular competitive exercise, the user respectively learner needs the description of the task, source data and, possibly, tools to employ.

Figure 6 is displaying the opportunities of downloading the training data file, of getting access to professional data mining software and, last but not least, to submit the solution in form of a PMML file.

PMML stands for predictive model markup language which is the XML standard for representing and communicating data mining results.

We complete this chapter by a discussion of how learners can experience true data mining by attacking the problem of e-mail spam filtering.

The available training datasets has 8000 e-mail messages some part of which (39%) are qualified as spam (positive examples) and the rest (61%) as not spam (negative examples). The data set is described by 833 attributes, including 832 binary ones and one numeric one. The only numeric attribute (ID) reflect the (unique) incoming number of e-mails within the company that provided the data.

First, there is the need to understand the problem and to become familiar with the data (cf. Figure 5). When a basic understanding is achieved, one needs to decide about the approach to undertake. Systems like QuDA usually offer some alternatives.

Among the algorithms implemented in QuDA, there are ID3 and its advancement C4.5 available.

Figure 7: Refreshing Knowledge about C4.5

Users may return to the DaMiT system to refresh their knowledge about C4.5 as illustrated here. They may study the fundamentals, comprehend examples and case studies or just go through the details of the algorithm to get an impression of its suitability.

Figure 8: Inspecting and Evaluating a Decision Tree

Figure 8 is displaying a QuDA window where a user can inspect the decision tree generated by C4.5. One may activate nodes of the tree and check how and why they work for filtering e-mail spam. Moreover, one can save his model in PMML for usage in other systems or for upload to the DaMiT system.

Note that with the knowledge and techniques of DaMiT, you can drive spam down to under 1%.

There should be a message shining through our detailed discussion of learning by doing when using the DaMiT system for educational purposes: This all may be relevant in application environments as well where understanding data and preparing decisions is essential. The stage is set for learning when doing.
5 BACKGROUND KNOWLEDGE FOR DECISION SUPPORT

Decision making requires to be knowledgeable on the outcome of solutions and to forecast or estimate, at least, the impact of the decision. Such a task is quite difficult to solve, in general. There is a saying accredited to Niels Bohr: “It’s difficult to predict, especially the future.”

A simpler task is the support of the generation of acceptable solutions for data already given. Such a task can be solved if the user knows how algorithms employed for solution generation do work on the data, which problems may arise from partially incorrect solutions and how generated solutions are statistically backed up.

Decisions clearly need to have a goal, but this is beyond the limits of the present paper. The authors assume that speaking about decisions and decision support one can reasonably assume some purpose.

Similarly, decisions are decisions among alternatives. Those are assumed as well and we only deal with the problem of supporting the choice between alternatives by means of knowledge derived from given data.

The crux is usually, as already said before, that the knowledge is not sitting in the data and hiding from the user’s view. Instead, there are lots of interpretations, and it is difficult to foresee which might be useful in the future.

For instance, one may understand why a majority of customers decided to cancel a contract in the past, but one has no guarantee that these reasons are really the dominating ones in the future. Or one may detect what characterizes a majority of incoming spam e-mails at present. But there is no guarantee that the senders are not changing their stuff. These are only a few examples to illustrate the unavoidable vagueness when exploring complex data and coming up with hopefully useful insights.

The most important consequence is: Whatever is derived from complex data to be used in the future, it is hypothetical by its very nature. And so are any decisions based on it.

There are several conclusions to be taken into account. Perhaps, the most important one is that users should never take results for granted. But how to avoid misinterpretations, if data mining results, for instance, are coming that clearly and understandable away like rules, e.g., to which the user can agree? The answer is that users must know about the procedures, the assumptions, the guesses and the parameters adjusted which all stay behind a particular result. This requires a deeper understanding than one can usually assume.

We will try to illustrate below how to proceed.

6 LEARNING ON DEMAND

When in practice problems do arise which may be explained or interpreted over large and usually distributed data, the essence is rarely sufficiently understood in the very beginning. Symptoms are recognized, but a useful diagnosis may take some time and may be laborious.

For illustration, an enterprise’s management may recognize a growing number of customers cancelling their business relations with that enterprise. A first self-evident management decision might be to ask somebody to look into the individual data and find out the reasons. If this fails, what to do next?

Even if no pressing problems urge management to inspect larger data bases, certain desires for cost reduction may lead to the wish of understanding relations not understood so far. For instance, if a mailing action in direct marketing shall be more focussed than it used to be before, one should find out which customers are very likely to respond and which are not. Again, a self-evident management decision might be to ask somebody to look into the data and tell which customers are to be addressed. If this fails, what to do next?

To have an appropriate e-learning system at the management’s fingertips may help a lot. You can get consulting about the general problem you are facing, you can get knowledge about approaches and technologies, you can get tools for attacking your problem, and, finally, you can get support in evaluating your own solution to the problem you have.

In a system like DaMiT, as seen above, you can find problems similar to the one you are facing. And you can get all this for free, because a large amount of the e-learning content is open to the public.

Figure 9: A Case in Direct Marketing Optimization

More generally speaking, with a system like DaMiT one can get consulting about the characteristics of a problem, about basic variants and crucial details to be considered, and about ways how to go...
forward towards a solution. This means already learning when doing.

Assume that a problem like that of finding those customers which are likely to respond to a mailing activity is understood as a classification problem. Let us further assume that the general principles of decision tree induction are understood and believed to be helpful. (If not, consult the system and learn more about this area.) Then it is a management decision to go for generating a decision tree classifier over the own data base.

In that case, data understanding and data preparation are inevitable steps. There is no hope at all to take your data as they are and start learning any useful classifier. In practice, this problem is generally awkwardly underestimated. In an enterprise, one may study the lessons and try the tools for data understanding (cf. section 4.1 above). Doing so means learning when doing.

If the tool has been chosen – we take QuDA in the sequel, for illustration – and the data are prepared, one can get involved in the laborious process of interactively generating a classifier.

Normally, one comes up with a first classifier, does inspect it and returns to the generation process.

Recall that data mining is both an art and a science, and whatever we generate by means of data mining technologies and tools, we only do arrive at hypotheses. There can not be any guarantee at all that models (like decision trees, e.g.) generated over given data behave as successful as expected over other data in the future.

There is a need to verify generated models. How to do that appropriately, which alternatives do exist, and what the results mean and how they are backed up by statistics, can be studied in DaMiT – another case of learning when doing.

Figure 10: QuDA in doing Decision Tree Induction

![Figure 10](image)

This figure is displaying the generation and inspection of a decision tree by means of QuDA over the data of a realistic direct marketing case study. There is a node of the decision tree highlighted and all customers classified by this node are listed in the window below.

A user should check whether he agrees with an approximate classification like this. If not, he has to return to the tree generation process.

Data mining tools offer different ways to take a subtree of the classifier generated so far and continue model generation at the point picked up.

Following the exemplified procedure in an enterprise when dealing with the own problem on the own data is not only the right way to solve a problem, it is also an instance of learning when doing.

Figure 11: Verifying a Generated Decision Tree

![Figure 11](image)

The present Figure 11 shows a verification result for the decision tree of Figure 10 generated by means of the C4.5 implementation of QuDA.

Once the building of a model has been completed, as shown in Figure 12, it can be exported from the generation tool and saved for use in the future.

Figure 12: A Solution to the Application Problem

![Figure 12](image)

XML standards like PMML allow for an integration into an enterprise’s IT infrastructure.

There is no closing sentence about learning when doing, because in areas like knowledge discovery and data mining, learning never ends. Systems like DaMiT are further developed to support this type of life-long learning.
7 SUMMARY & CONCLUSIONS

On the one hand, the Internet in pervading the world has changed our daily life considerably, and it is currently changing human learning at all stages ranging from ordinary schools through higher education to professional training and life-long learning. Technology enhanced learning is providing quite new opportunities.

On the other hand, there is the obviously eternal gap between academia and practice which appears as a certain reluctance to e-learning in practice.

Despite these obvious difficulties, we are at the cusp of closing the gap. As exemplified in the data mining domain, even academic e-learning has the urgent need of doing, i.e. learning by doing. In practice, for sure, there is the need of knowledgeable doing which leads to learning when doing. Enterprise application integration (Linthicum, 2000) will allow for a proper dovetailing of learning and doing. Data mining may be an area worth for doing it now.

8 ACKNOWLEDGEMENTS

The present paper is based on work of colleagues and friends from 11 academic institutions and draws benefit from many other partners using the DaMIT system in their educational practice.

Basic work on the paper has been done when the first four authors met for a conference in Sapporo, Japan. These authors are grateful to those Japanese colleagues like Yuzuru Tanaka, Akihiro Yamamoto, Makoto Haraguchi and Yoshiaki Okubo who made this meeting possible.

The paper has been completed during the first author’s stay as a visiting professor at the Meme Media Lab. of Hokkaido University, Sapporo, Japan.

REFERENCES


