

PROJECT MLEXAI: APPLYING MACHINE LEARNING TO WEB DOCUMENT CLASSIFICATION*

Ingrid Russell

*University of Hartford, West Hartford, CT
Phone: 01-860-768-419, Email: irussell@hartford.edu*

Zdravko Markov

*Central Connecticut State University, New Britain, CT
Phone: 01-860-832-2712, Email: markovZ@ccsu.edu*

Susan Coleman

*University of Hartford, West Hartford, CT, USA
Phone: 01-860-768-4690, Email: scoleman@hartford.edu*

ABSTRACT

We present work on project MLE_xAI, funded by the National Science Foundation with a goal of unifying the artificial intelligence (AI) course around the theme of machine learning. Our work involves the development, implementation, and testing of an adaptable framework for the presentation of core AI topics that emphasizes the relationship between AI and computer science. A suite of adaptable hands-on laboratory projects that can be closely integrated into a one-term AI course and which would supplement introductory AI texts has been developed. The paper focuses on one of these projects, how it meets our goal, and presents our experiences using it. The project involves the development of a learning system for web document classification. Students investigate the process of classifying hypertext documents, called tagging, and apply machine learning techniques and data mining tools for automatic tagging. A summary of our experiences using the projects during four course offerings over the last two years are also presented.

* Copyright © 2007 by the Consortium for Computing Sciences in Colleges. Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the CCSC copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Consortium for Computing Sciences in Colleges. To copy otherwise, or to republish, requires a fee and/or specific permission.

1. INTRODUCTION

An introductory Artificial Intelligence (AI) course provides students with basic knowledge of the theory and practice of AI as a discipline concerned with the methodology and technology for solving problems that are difficult to solve by other means. It is generally recognized that an introductory Artificial Intelligence course is challenging to teach. This is, in part, due to the diverse and seemingly disconnected core AI topics that are typically covered. Recently, work has been done to address the diversity of topics covered in the course and to create a theme-based approach. Russell and Norvig present an agent-centered approach [9]. A number of institutions have been working to integrate Robotics into the AI course [1,2,3]. Our work incorporates machine learning as a unifying theme for the AI course. Machine learning is a sub-field within the field of artificial intelligence that is concerned with building computer systems that have the ability to improve their performance in a given domain through experience. Machine learning is inherently connected with the AI core topics and provides methodology and technology to enhance real-world applications within many of these topics. Machine learning also provides a bridge between AI technology and modern software engineering. In his article, Mitchell discusses the increasingly important role that machine learning plays in the software world and identifies three important areas: data mining, difficult-to-program applications, and customized software applications [6].

Project MLE_xAI involves the development of a suite of adaptable, hands-on projects that can be closely integrated into a one-term AI course and which would supplement introductory AI texts. Each project involves the design and implementation of a learning system which will enhance a particular commonly-deployed application. The goal is to enhance the student learning experience in the introductory artificial intelligence course. The focus of this paper is on a project that involves the development of a learning system for web document classification. Students investigate the process of tagging web pages using the Open Directory structure and apply machine learning techniques for automatic tagging. Our experiences using this material at two institutions, University of Hartford and Central Connecticut State University, during four course offerings over the last two years will also be presented.

2. OVERVIEW OF PROJECT MLEXAI

The AI course at both institutions involved is offered at the junior and senior levels with data structures as a pre-requisite. Typically, this course provides students with basic knowledge of the theory and practice of artificial intelligence as a discipline concerning intelligent agents capable of deciding what to do and doing it. Our offerings had been largely consistent with traditional offerings in providing students with basic knowledge of the theory and practice of AI as a discipline. The course covers core AI topics such as search algorithms, knowledge representation, and reasoning. The objectives of our introductory AI course are:

- To have an appreciation for and understanding of both the achievements of AI and the theory underlying those achievements.
- To have an appreciation for the engineering issues underlying the design of AI systems.

- To have an understanding of the basic issues of knowledge representation and blind and heuristic search, as well as an understanding of other topics such as minimax, resolution, etc. that play an important role in AI programs.
- To have a basic understanding of some of the more advanced topics of AI such as learning.

Traditionally, in addition to the core topics mentioned above, brief introductions to several sub-fields of AI have been provided. However, we believe that presenting an overview of many sub-fields provides students with superficial understanding of these areas. Instead, our approach would cover fewer sub-fields, concentrating on machine learning and what is needed to support the projects being proposed. The difficulties mentioned above associated with the introductory AI course, combined with the increasingly important role of machine learning in computer science, are the motivating factors for this work. The goal is to develop a framework for teaching core AI topics with a unifying theme of machine learning and to enhance the student learning experience in the introductory artificial intelligence course by:

- (1) Introducing machine learning elements into the AI course.
- (2) Implementing a set of unifying machine learning laboratory projects to tie together the core AI topics.
- (3) Developing, applying, and testing an adaptable framework for the presentation of core AI topics which emphasizes the important relationship between AI and computer science in general, and software development in particular.

These objectives were accomplished through the development, implementation, and testing of a suite of adaptable and self-contained, hands-on open laboratory projects that can be closely integrated into the AI course. We have developed six such projects. Each of our projects is intended to be a term-long project with several deliverables throughout the term and will involve the development of learning systems in commonly deployed applications. The projects allow for varying levels of programming sophistication, with some allowing for more focus on tools, while others have significant programming components. In addition, our framework is adaptable to allow instructors to extend it based on local needs. Details on the projects are available at: <http://uhaweb.hartford.edu/compsci/ccli>. An overview of Project MLExAI and samples of other course materials developed under this grant are published in [4, 5, 7, 8, 10, 11]. A detailed description of the Web Document Classification project and how it meets our objectives is presented next.

3. SAMPLE PROJECT: WEB DOCUMENT CLASSIFICATION

Along with search engines, topic directories are the most popular sites on the Web. Topic directories organize web pages in a hierarchical structure (taxonomy or ontology) according to their content. The purpose of this structuring is twofold: first, it helps web searches focus on the relevant collection of Web documents. The ultimate goal here is to organize the entire web into a directory, where each web page has its place in the hierarchy and thus can be easily identified and accessed. The Open Directory Project (dmoz.org) and About.com are some of the best-known projects in this area. Second, the topic directories can be used to classify web pages or associate them with known topics.

This process is called tagging and can be used to extend the directories themselves. The aim of the web document classification project is to investigate the process of tagging web pages using the Open Directory structure and to apply Machine Learning techniques for automatic tagging. This would help to filter out the responses of a search engine or to rank them according to their relevance to a topic specified by the user. The project covers important AI topics such as search and knowledge representation. Students work on a very small section of dmoz. Specifically, they select 5 topics from dmoz and create a system that will automate adding web pages to that branch of dmoz. The plan is to use machine learning to identify which subtree of dmoz a web page belongs to. Students select approximately 100 web documents which they know where they belong in dmoz, train/teach the system to recognize how to classify these web documents, and then use it to categorize new web documents, i.e., identify which subdirectory of dmoz the new document should be added to. In the following section, we present details of the project along with student deliverables at each phase.

The project is split into three major parts, which are also phases in the overall process of knowledge extraction from the web and classification of web documents (tagging). As this process is interactive and iterative in nature, the phases may be included in a loop structure that would allow each phase to be revisited so that some feedback from latter phases can be used. The parts are well defined and can be developed separately (e.g. by different teams) and then put together as components in a semi-automated system or executed manually. Hereafter we describe the project phases along with the deliverables that the students need to submit on completion of each phase.

3.1 Web Document Collection

The purpose of this phase is to collect sets of web documents belonging to different topics (subject area). The basic idea is to use a topic directory structure. Such structures are available from dmoz.org (the Open Directory project), the yahoo directory (dir.yahoo.com), about.com and many other web sites that provide access to web pages grouped by topic or subject. These topic structures have to be examined in order to find several topics (e.g. 5), each of which is well represented by a set of documents (at least 10). Alternative approaches could be extracting web documents manually from the list of hits returned by a search engine using a general keyword search or collecting web pages from the web page structure of a large organization.

Deliverable: The outcome of this phase is a collection of several sets of web documents representing different topics or subjects, where the following restrictions apply:

a) As these topics will be used for learning and classification experiments at later phases they have to form a specific structure (part of the topic hierarchy). It is good to have topics at different levels of the topic hierarchy and with different distances between them (a distance between two topics can be defined as the number of predecessors to the first common parent in the hierarchy). It is good to have topics at different levels of the topic hierarchy and with different distances between them (a distance between two topics can be defined as the number of predecessors to the first common parent in the hierarchy). An example of such structure is:

topic1 > topic2 > topic3 topic1 > topic2 > topic4
 topic1 > topic5 > topic6 topic1 > topic7 > topic8
 topic1 > topic9

The set of topics here is {topic3, topic4, topic6, topic8, topic9}. Also, it would be interesting to find topics which are subtopics of two different topics. An example of this is:

Top > ... > topic2 > topic4 Top > ... > topic5 > topic4

b) There must be at least 5 different topics with at least 20 documents in each.

c) Each document should contain a minimum amount of text. This may be measured with the number of words (excluding articles and punctuation marks).

d) Each document should be in HTML format and contain HTML tags as title, headings or font modifiers.

Figure 1 shows a student sample structure and pages taken from dmoz.org site for phase 1 of the project.

Topic 1	Topic 2	Topic 3
Top: <u>Computers</u>: <u>Artificial Intelligence</u>: Machine Learning		
Topic 1	Topic 2	Topic 4
Top: <u>Computers</u>: <u>Artificial Intelligence</u>: Agents		
Topic 1	Topic 5	Topic 6
Top: <u>Computers</u>: <u>Algorithms</u>: Sorting and		
Topic 1	Topic 7	Topic 8
Top: <u>Computers</u>: <u>Multimedia</u>: MPEG		
Topic 1	Topic 9	
<u>Top: <u>Computers</u>: History</u>		

Figure 1: Phase 1 Sample Structure

3.2 Feature Extraction and Data Preparation

At this phase the web documents are represented by feature vectors, which in turn are used to form a training data set for the Machine Learning phase. Three basic steps are involved. First, students select a number of terms (words) whose presence or absence in each document can be used to describe the document topic. This can be done manually by using some domain expertise for each topic or automatically by using a statistical text processing system. The latter is based on putting all documents together and sorting in ascending order all words appearing in all documents by their frequency. The first N words in the sorted sequence can be used to represent the documents with vectors of size N. Using the selected set of terms as features (attributes), students create a feature vector (tuple) for each document with Boolean values corresponding to each attribute (1 if the term is in the document, 0 if it is not). A more sophisticated approach for determining the attributes values can be used too. It is based on using the term frequencies scaled in some way to normalize the document length. Further, the HTML tags may be used to modify

the attribute values of the terms appearing with the scope of some tags (for example, increase the values for titles, headings and emphasized terms).

In the final step of this phase, students create a data set in the ARFF (Attribute-Relation File Format) format to be used by the Weka Machine Learning system [12]. An ARFF file is a text file which defines the attribute types (for the Boolean values they will be nominal, and for the frequency-based ones numeric) and lists all document feature vectors along with their class value (the document topic).

Deliverable: Students submit ARFF data files containing the feature vectors for all web documents collected in phase 1. It is recommended that students prepare several files by using different approaches to feature extraction. For example, one with Boolean attributes, one with numeric based on text only, and one with numeric using the html information. Versions of the data sets with different number of attributes can also be prepared. The idea of preparing all these data sets is twofold. First, by experimenting with different data sets and different machine learning algorithms, the best classification model can be found. Second, by evaluating all these models, students will understand the importance of various parameters of the input data for the quality of learning and classification.

3.3 Machine Learning Phase

At this phase, machine learning algorithms are used to create models of the data sets. These models are then used for two purposes. First, the accuracy of the initial topic structure is evaluated and second, new web documents are classified into existing topics. For both purposes we use the Weka Data Mining System – a free Machine Learning software package [12]. This is one of the most popular machine learning systems used for educational purposes. It is the companion software package of a book on machine learning and Data Mining [13]. An online tutorial allows students to quickly become familiar with the functionality of the system. Students explore several machine learning algorithms and their performance in automatic tagging of the web documents. The machine learning phase of the project consists of the following steps. Students load the ARFF files created in phase 2 and use Weka to explore several machine learning algorithms and their performance in automatic tagging of the web documents. For example, students use Weka's decision tree algorithm (J48) to examine the decision trees generated with different data sets. Students are asked to find the most important terms for each data set (the terms appearing on the top of the tree) and how they change with changing the data set. By checking the classification accuracy and the confusion matrix obtained with 10-fold cross validation, they find out which topic is best represented by the decision tree. The Cross-Validation method randomly reorders the dataset and then splits the dataset into 'n' folds of equal size. For each iteration, one fold is used for testing and the other n-1 folds are used for training.

Students are also asked to use the Naïve Bayes and K-Nearest Neighbor algorithms and compare their classification accuracy and confusion matrices obtained with 10-fold cross validation with the ones produced by the decision tree. The last step is to classify new web documents. The process involves selecting web documents from the same subject areas, but not belonging to the original set of documents prepared in phase 1.

Documents from different topics are also generated. Students then apply feature extraction and create ARFF files, each representing one document. Using the Weka test set option, a classification of the new documents is accomplished. A comparison of the original topic with the one predicted by Weka is done. Figure 2 presents a student sample partial decision tree constructed by the J48 classifier which indicates how the classifier uses the attributes to make a decision. The leaf nodes indicate which class a document will be assigned to, should that node be reached. The numbers in parentheses after the leaf nodes indicate the number of documents assigned to that node, followed by the number of incorrectly classified documents. Figure 3 shows results of testing using the three classification algorithms and 10-fold cross validation. This is a partial output of the system that includes the most important information.

Machine Learning Phase Deliverable: Students are asked to explain the decision tree learning algorithm in terms of state space search. This phase of the project requires writing a report on the experiments performed. The report includes a description of the experiments (inputs data, Weka outputs), and interpretation and analysis of the results with respect to the original problem stated in the project, document classification.

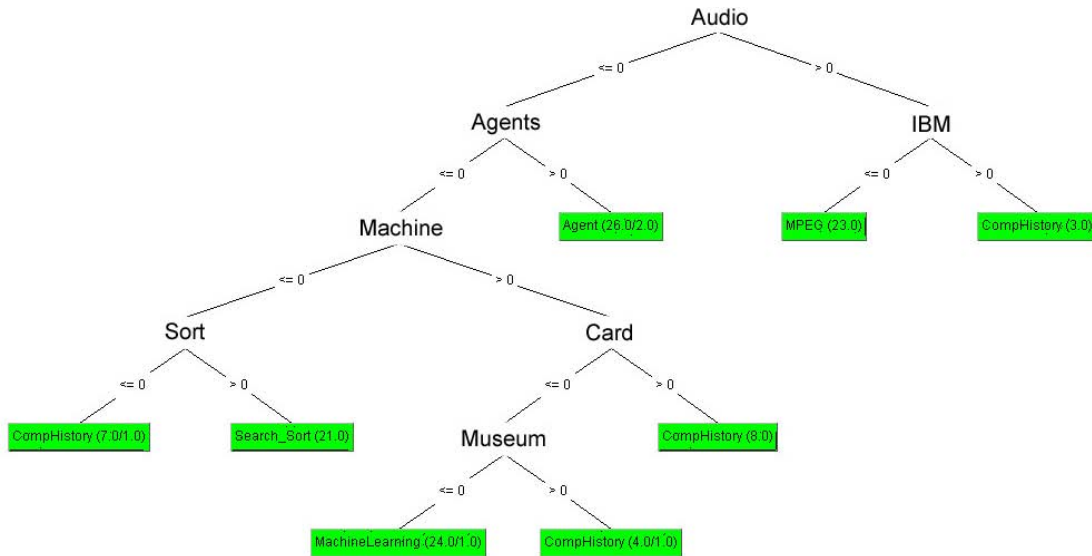


Figure 2: Partial Decision Tree constructed by the J48 Classifier

==== Stratified Cross-Validation =====	
Scheme	J48
Correctly Classified Inst.	(101) 87.069 %
Incorrectly Classified Inst.	(15) 12.931 %
Total Number of Instances	116
==== Stratified Cross-Validation =====	
Scheme	K-NN
Correctly Classified Inst.	(107) 92.2414 %
Incorrectly Classified Inst.	(9) 7.7586 %

Total Number of Instances	116
===== Stratified Cross-Validation =====	
Scheme: Naïve Bayes	
Correctly Classified Inst.(112)	96.5517 %
Incorrectly Classified Inst.(4)	3.4483 %
Total Number of Instances	116

Figure 3: Results of Testing

4. EXPERIENCES

The Web Document Classification project combines a number of important AI areas. Web search engines use advanced search algorithms and information retrieval techniques to find web pages. In addition, they use knowledge representation techniques to organize and structure the search results and create topic directories. While enforcing traditional AI core topics, using a unified example, in this case web document classification, the project allowed the discussion of various issues related to machine learning including:

- The basic concepts and techniques of machine learning.
- Issues involved in the implementation of a learning system.
- The role of learning in improved performance and in allowing a system to adapt based on previous experiences.
- The important role data preparation and feature extraction play in machine learning.
- The vector space model for representing web documents and a variety of feature extraction techniques combined with the pros and cons of each in identifying and classifying documents by feature vectors.
- The importance of model evaluation in machine learning and in particular the training and testing framework used to choose the best model for web page classification.

The projects were implemented and tested at the University of Hartford during fall 2004 and spring 2006 and at the Central Connecticut State University during spring 2005 and spring 2006. Led by the project evaluator, an independent evaluation of this work involved students, faculty, internal and external evaluators, and an external advisory board. Feedback from students during the first offerings helped in the refinement of the projects and in their use in later offerings of the courses. Students were asked to complete a survey comprised of 22 ranked questions and two open-ended questions. In addition, a sample of students was interviewed from each location using either a direct (group) interview or a group conference call.

Table 1: Student Survey Results

Question	%A & SA
The student project contributed to my overall understanding of the material in the course.	92-100%
After taking this course I feel that I have a good understanding of the fundamental concepts in AI.	92-100%
Based on my experience with this course, I would like to learn more about the field of AI.	80-100%
I had a positive learning experience in this course.	87-100%

We include in Table 1 the range of percentages over all evaluation locations and terms of those who replied with ‘Agree’ or ‘Strongly Agree’ to 4 selected questions. Students expressed their enthusiasm in a variety of ways. They felt that the project was interesting to work on and contributed to their overall understanding of the material in the course. By the end of the semester, most felt that they had a good understanding of the fundamentals of AI and Machine Learning. The majority of the students surveyed indicated that they would like to learn more about both AI and Machine Learning. They also indicated that they would like further opportunities to apply the problem-solving techniques in the future. The majority of students was enthusiastic about the projects, and felt that they were an effective way to introduce Machine Learning. Their responses indicated a high level of confidence in their ability to apply these problem-solving techniques in other situations. In summarizing their experience, students stated that they had a very positive learning experience in the course. These sentiments were reinforced by their responses to an open-ended question that asked what they liked best about the course. The group interviews and conference calls provided further elaboration for responses provided in the written evaluation.

Students were very satisfied with the organization of the course, the number of topics covered, and the hands-on nature of the team project. They liked being able to use AI principles to solve real problems. In fact, the majority stated that this was the most positive aspect of the course. Further, most students indicated that the projects enhanced their understanding and gave them new ways to think about the problems. They felt they had a much better understanding of the importance of AI and its broad applicability. Students indicated a high level of satisfaction with the course and with the method of instruction. Many also indicated the desire to pursue additional study or research in AI and Machine Learning. Following the courses, several students enrolled in independent study courses and pursued research projects.

5. CONCLUSION

We presented work on a project whose goal is to develop a framework for teaching core AI topics with a unifying theme of machine learning. A suite of hands-on laboratory projects was developed that can be integrated into a one-term AI course. We presented a sample project related to Web Document Classification. The goal of this project is to investigate the process of tagging web pages using the topic directory structures and to apply Machine Learning techniques for automatic tagging. This would help in filtering out the responses of a search engine or ranking them according to their relevance to a topic specified by the user. Assessment of our work and the projects demonstrated that student experiences were very positive. The results indicate that the projects enhanced the student learning experience in the introductory artificial intelligence course. While working on this project, students learned the basics of Information Retrieval, Data Mining and Machine Learning, and gained experience in using recent software applications in these areas. Most importantly, they demonstrated a better understanding of fundamental AI concepts such as Knowledge Representation and Search, which play important roles in the areas mentioned above.

6. ACKNOWLEDGEMENTS

This work is supported in part by NSF grant DUE CCLI-A&I Award Number 0409497.

7. REFERENCES

- [1] Dodds, Z. et al. (eds) *Robots and Robotics in Undergraduate AI Education*, Special Issue of AI Magazine, 27(1), AAAI Press, 2006.
- [2] Kumar, A., Kumar, D, Russell, I., “Non-Traditional Projects in the Undergraduate AI Course”, Proceedings of the 37th Annual SIGCSE Technical Symposium on Computer Science Education, ACM Press, March 2006.
- [3] Kumar, D., and Meeden, L., A Robot Laboratory for Teaching Artificial Intelligence, *Proceedings of SIGCSE*, ACM Press, New York, NY, 1998, pp.341-344.
- [4] Markov, Z., Russell, I., Neller, T., and Zlatareva, N., “Pedagogical Possibilities for the N-Puzzle Problem”, *Proceedings of The Frontiers in Education Conference*, IEEE Press, November 2006.
- [5] Markov, Z., Russell, I., Neller, T., and Coleman, S., “Enhancing Undergraduate AI Courses through Machine Learning Projects”, *Proceedings of the Frontiers in Education Conference*, IEEE Press, October 2005.
- [6] Mitchell, T., Does Machine Learning Really Work, *AI Magazine*, Vol. 18, No. 3, AAAI Press, Fall 1997.
- [7] Neller, T., Russell, I., Presser, C., and Markov, Z., “Pedagogical Possibilities for the Dice Game Pig”, *Journal of Computing Sciences in Colleges*, 21(5), April 2006.
- [8] Neller, T., Markov, Z., and Russell, I., “Clue Deduction: Professor Plum Teaches Logic”, *Proceedings of the 19th International FLAIRS AI Conference*, AAAI Press, May 2006.
- [9] Russell, S. J. and Norvig, P., *Artificial Intelligence: A Modern Approach*, Upper Saddle River, NJ: Prentice-Hall, Second edition, 2002.
- [10] Russell, I., Markov, Z., Neller, T., Georgiopoulos, M., and Coleman, S., Unifying Undergraduate AI Courses through Machine Learning projects, *Proceedings of the American Society for Engineering Education Conference*, June 2005.
- [11] Russell, I., Markov, Z., and Neller, T., “Teaching AI through Machine Learning Projects”, *Proceedings of the 11th Annual Conference on Innovation and Technology in Computer Science Education*, ACM Press, June 2006.
- [12] *Weka Home Page*, <http://www.cs.waikato.ac.nz/~ml/weka>. [13] Witten, I. H. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* (Second Edition), Morgan Kaufmann, 2005.