Automated coding of AD-417 forms using case-based reasoning
by Venkatesh Thati

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science
Montana State University

Abstract:
The AD-417 forms are used by researchers to submit their proposals to the United States Department of Agriculture. The researchers are required to enter codes belonging to various categories, that are pertinent to their project, in the AD-417 form. The Classifier system has been developed to reduce the time and increase the accuracy in filling out these forms. The basis for this system, a machine learning technique called Case-Based Reasoning, is described along with a discussion of its advantages and disadvantages. The components of the Classifier are described and results that indicate its performance are provided.

AUTOMATED CODING OF AD-417 FORMS USING CASE-BASED REASONING

by

Venkatesh Thati

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Master of Science

in

Computer Science

MONTANA STATE UNIVERSITY
Bozeman, Montana

January, 1995

APPROVAL

of a thesis submitted by

Venkatesh Thati

This thesis has been read by each member of the thesis committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

1/19/95
Date

John T. Paton
Chairperson, Graduate Committee

Approved for the Major Department

1/19/95
Date

J Denbigh Starkey
Head, Major Department

Approved for the College of Graduate Studies

1/25/95
Date

R L Brown
Graduate Dean

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library.

If I have indicated my intention to copyright this thesis by including a copyright notice page, copying is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for permission for extended quotation from or reproduction of this thesis in whole or in parts may be granted only by the copyright holder.

Signature T. VanLanteh
Date January 19, 1995

TABLE OF CONTENTS

TABLE OF CONTENTS - Continued

Page

List of Figures

ABSTRACT

The AD-417 forms are used by researchers to submit their proposals to the United States Department of Agriculture. The researchers are required to enter codes belonging to various categories, that are pertinent to their project, in the AD-417 form. The Classifier system has been developed to reduce the time and increase the accuracy in filling out these forms. The basis for this system, a machine learning technique called Case-Based Reasoning, is described along with a discussion of its advantages and disadvantages. The components of the Classifier are described and results that indicate its performance are provided.

Case-Based Reasoning

Case-based reasoning is a technique which attempts to model the concept of experience. It is based on the psychological theories on how experience contributes towards understanding and solving problems.

From the perspective of case-based reasoning, the best person to solve a problem is not necessarily the smartest person, it is the person with the most experience. It considers a person who has solved a similar problem before to be the one best able to solve the problem now. If we had two computer systems to solve a certain kind of problem, and if we could make one of them acquire "experience" as it goes about the process of solving problems, there is a good chance that given a new problem the system with experience will perform better than the other.

A case can be defined as a description of a problem, an attempt at a solution and an outcome of the effort[4]. Cases can be pretty complex, encapsulating many facts and relationships in a specific context. A case base would then be a collection of cases, i.e., an accumulated body of problem solving experiences. As the cases increase in number and diversity, so does the usefulness of the case base.

Case-based reasoning can be formally defined as a method of solving new problems by remembering old problems and adapting their solutions. It comprises a memory model for representing, indexing and organizing past cases and a process model for retrieving and modifying old cases and assimilating

new ones. This method combines reasoning with learning. It spans the whole reasoning cycle. A situation is encountered. Old situations are used to understand it. Then the new situation is inserted into the case base alongside the others to be used another time.

The key to the method of case-based reasoning is remembering[5]. Remembering has two parts,

- Integrating cases into memory when they happen
- Recalling the appropriate cases in later situations

This related set of issues is called the Indexing Problem. In broad terms, it means finding in the case base the case closest to a new one. In narrower terms one can think of it as a two-part problem,

- The first part involves assigning indices or labels to cases, when putting them into memory that describe the situations to which they are applicable, so that they can be recalled later.
- The second part involves elaborating a new situation in enough detail so that the indices it would have had if it were already in the case base are identified.

The technique of case-based reasoning is appropriate for many types of applications. Potential domains can be identified by the degree to which they meet the following criteria,

- First, when experts solve a problem in the domain if they refer to specific cases that helped illuminate, describe, classify or solve the current problem then case-based reasoning would be suitable for this domain.

- Second, a case-based reasoning system is only useful if similar problems are seen again and again. If each problem and its solution is unique there is little value in storing it.

- Third, case-based reasoning systems, neural networks and expert systems have somewhat overlapping capabilities. It is important to determine which is most appropriate. Case-based reasoning is appropriate when there is little understanding of underlying causal relationships in the domain and when weak explanations characterize the selection of solutions.

- Fourth, cases should be relatively compact and be defined in a manner that makes the identification of indices which can be used for matching relatively simple. For example, historical cases of corporate development are probably too vaguely described. So they won't be of much use.

- Fifth, an application like a help desk is an ideal candidate for applying case-based reasoning because there are many people at work, turnover is high and expertise is distributed. On the other hand if we had a stable technology all the problems of which can be solved by a single person and if the continuing presence of this person can be relied on, then case-based reasoning would be less cost effective.

- Sixth, it should be cost-effective to obtain and store the cases.

Based on these criteria some of the potential areas to apply case-based reasoning are help desks, medical diagnosis, scheduling etc.

Advantages of Case-based reasoning

Some of the benefits of case-based reasoning are,

- A case base becomes useful with the first case. It is not necessary to wait until all the cases have been developed before the system can be used. As cases are added, the system becomes more useful.

- A case base captures knowledge easily. The structure of cases is much less constrained than rules are. There is no need for complex interrelations between cases as there are between rules. Consequently case bases come on-line faster and they stay on-line even as cases are being altered or eliminated. Learning is incremental.

- Case bases are more understandable. The basic organization and functioning of case-based reasoning systems is logical and easy to follow. People feel better about using systems that are more understandable.

- Case-based reasoning augments human capabilities. A case-based reasoning system can track more cases than a person can, and as with other computer systems it thoroughly and neutrally evaluates all possibilities before making a recommendation.

- Case-based reasoning systems facilitate the incorporation of new knowledge. New cases can be added rapidly to a case-based reasoning system, thereby increasing its usefulness.

Disadvantages of Case-based reasoning

Some of the drawbacks of case-based reasoning are,

- It is a form of supervised learning, in the sense that during training it requires a teacher in the form of an expert to enter in the correct solution when the system comes up with unsatisfactory solutions.
- Case-based reasoning does not actually find the reasons behind the relationships it reveals, it simply deals on a case-by-case level.
- As the case base grows, due to the large number of cases being processed certain inconsistencies may crop up in the case base which can easily go unnoticed.
- The process of situation assessment, that is the extent to which a reasoner can interpret a new case and determine which kinds of cases are most likely to be useful, needs to be further improved.