

Clearing the Metadata Hurdle for Inexperienced Users: A Case-Based Reasoning Approach

Edward Brent, Idea Works, Inc.
Albert Anderson, Public Data Queries, Inc.
Lisa Neidert, University of Michigan
Pawel Slusarz, Idea Works, Inc.

Acknowledgement

Public Data Queries, Inc., is developing PDQ-Explore and the PDQ-Expert interface with the support of the Small Business Innovation Research (SBIR) and Small Business Technology Transfer Research (STTR) programs of the National Institute of Child Health and Human Development (NICHD) and the National Institute on Aging (NIA) of the National Institutes of Health (NIH). This portion of the project is supported by NICHD award R43 HD37738.

Homepages

- www.ideaworks.com
- www.pdq.com
- www.psc.isr.umich.edu

The Metadata Hurdle

- For inexperienced users of complex data sets, mastering the metadata can be a formidable task:
 - Which data sets are relevant to a given concern?
 - How “good” are the items in terms of data quality as well as their availability for the time periods, geography, and populations of interest?
 - Are the items acceptable measures of more general concepts?
 - What are the nuances and “gotchas?”

Introduction

- Digital representations of data make it possible to have intelligent interactive social science data capable of helping users formulate questions, specify analyses, and interpret results.
- This presentation describes an intelligent user interface now under development for the PDQ-Explore information system that strives to achieve some of these objectives.
- The interface uses the Idea Works' Qualrus Intelligent Qualitative Analysis Program to imbed case-based reasoning within a system of logical rules and semantic networks.

PDQ-Explore

- The PDQ-Explore information system combines paralleled high performance processors, data cached in random access memory, and efficient retrieval algorithms capable of processing tens of millions of records per second.
- Complex queries can be defined and executed in real time to produce tabulations, summary statistics, correlation matrices, and data extracts.

A Demonstration Version

- PDQ-Explore is structured as a client-server architecture with a graphical user interface accessible over the World-Wide Web.
- A demonstration version of PDQ-Explore with a preliminary Web-based interface is accessible from the Public Data Queries, Inc. home page at www.pdq.com.
- Example queries and the graphical-based client program can also be downloaded from that site.

Case-Based Reasoning (CBR)

- Case-Based Reasoning (CBR) attempts to solve new problems by making an analogy to a previous problem and then adapting its solution to fit the current problem (Kolodner and Leake, 1996).

CBR is Pervasive

- CBR is commonly used in many settings, including:
 - physicians diagnosing an illness based on previous cases;
 - lawyers arguing their case based on previous court cases;
 - real estate appraisers assessing the value of a house based on recent sales of similar houses.

Four Basic Steps in CBR

- Retrieve similar cases;
- Map the solution from the previous problem to the current one;
- Revise the solution to fit what is different about the current problem;
- Save the results as a new case to help with future problems.

Example: Real-Estate Appraisal

- These four steps can be illustrated for an example of case-based reasoning, real-estate appraisal:
 - 1. Examine recently sold houses similar in location, size, features, and condition;
 - 2. Select the three most similar houses and show their size and features side by side;

Real Estate Appraisal (continued)

- 3. Beginning with the price and features of each house, adjust for differences in size and features to estimate the cost of the current house;
 - E.g., use a common square-foot cost of construction to adjust for differences in square-feet.
- 4. Once the new house is sold, save it as an additional case for future appraisals.

Indexing Cases

- Each case consists of both the problem and its solution.
- Cases are assigned a set of indices representing key features used to identify similar cases.
- Indices here are key variables from the Census data along with other features of the query.

Retrieving Similar Cases

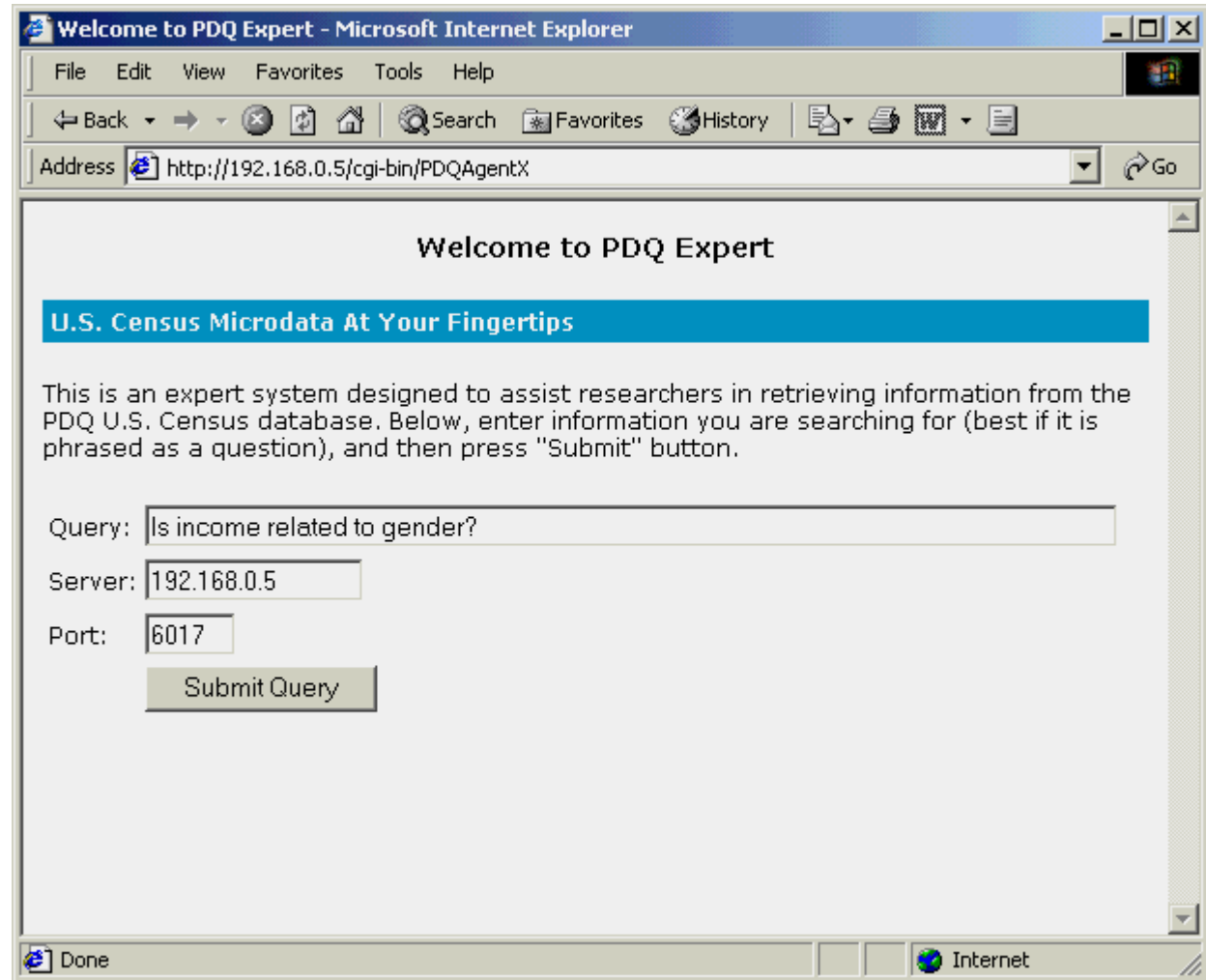
- Each old case is compared to the current case by computing a similarity metric (Golding and Rosenbloom, 1993)
- The ten most similar existing cases are displayed.

Adapting Cases to Fit the Current Problem

- Retrieved cases similar to the current problem must be adapted to fit the current problem if they are selected.
- Currently, users make this adaptation themselves.
- However, we hope in the future to automate at least some aspects of this adaptation.

The PDQ-Expert WWW Interface

- The WWW Interface for PDQ-Expert lets users type in a free-form question based on U.S. Census Microdata (IPUMS).
- Users enter their question in the “Query” field then click on the “Submit Query” button.



The System's Understanding of Your Query

- The first part of PDQ-Expert's response to the user is a restatement of the user's question as understood by the system.
- The purpose of this is to help the user see what the program thinks they are asking so they can identify any areas where the program may be going astray.

PDQ Interaction - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites History Print

Address <http://192.168.0.5/cgi-bin/PDQAgentX> Go

PDQ Expert Query Interaction

U.S. Census Microdata At Your Fingertips

[New](#) [Continue](#) [Similar](#) [Feedback](#)

System's understanding of your query:

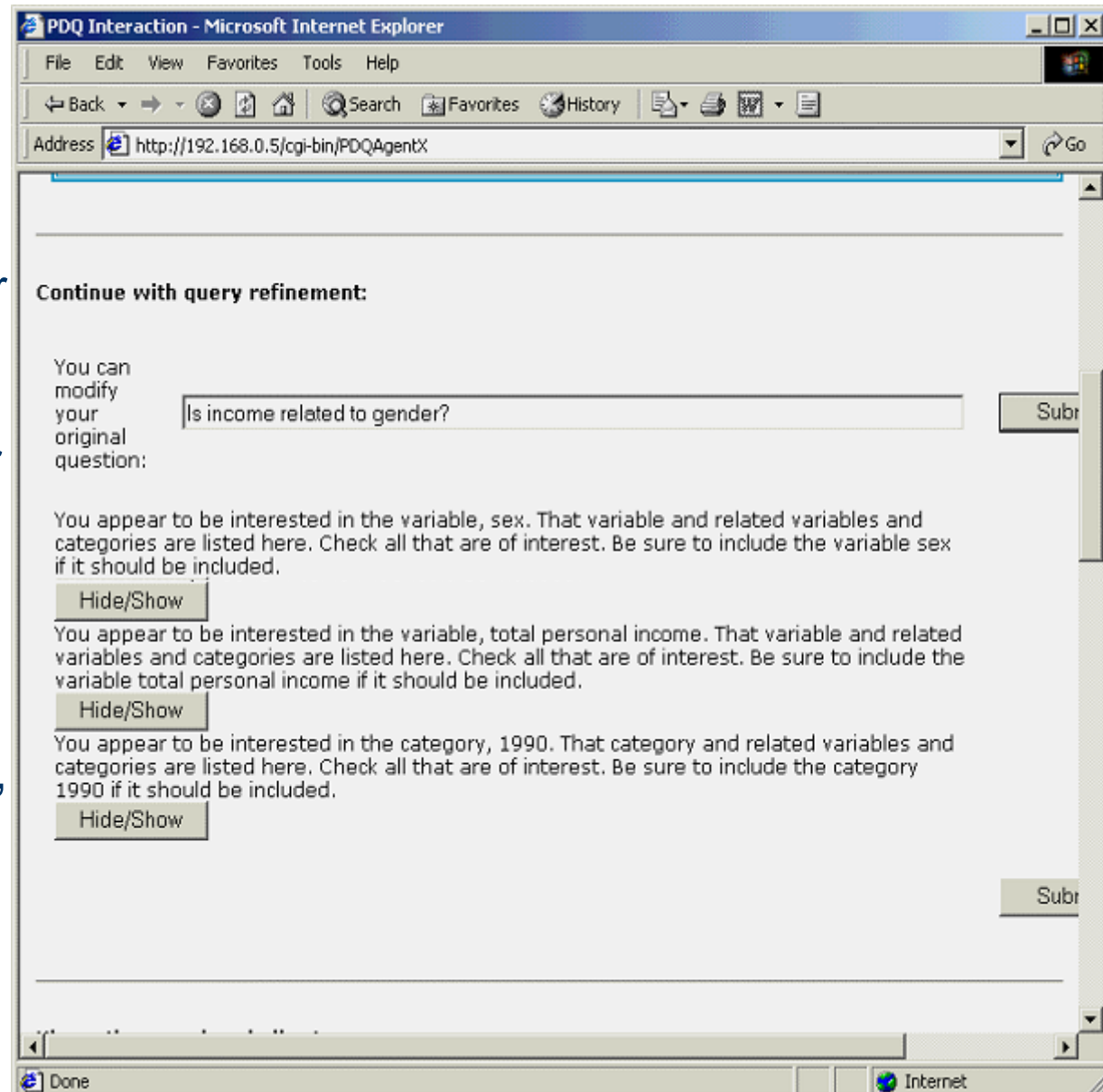
Latest PDQ Expert Version 2:12PM CDT May 29, 2002
Query: Is income related to gender?
Data Set: IPUMS Concatenated
Timeframe: 1990
Sex is directly measured by the variable, SEX(IPUMS) available in the IPUMS data set for the year, 1990.
Total personal income is directly measured by the variable, INCTOT(1990) available in the IPUMS data set for the year, 1990.
1990 is directly measured by the category, YEAR=99, of the variable, YEAR (PDQ IPUMS RECODE) available in the IPUMS data set for the year, 1990.
(Units are persons.)
Type: summary statistics
Weight: default
Expression: INCTOT(1950)
Universe: YEAR=99
(The universe is 1990.)
(The Query should examine a table relating sex to total personal income.)
Row: none
Column: sex
(There is no clear causal ordering for the variables, but perhaps sex is the independent variable and total personal income is the dependent variable in the table.)
CLARIFY CONCEPTS
Done.
Similar segments stored in file similarsegments.txt in the project folder.

Continue with query refinement:

Done Internet

Refining Your Query

- After reviewing what the program thinks the user was asking, the next step is to consider key concepts from their question they may want to change or clarify.
- For example, here the original question suggests the concepts, “sex,” “total personal income,” and “1990.”



Modifying the Query to Address Related Concepts

- Clicking the “Hide/Show” button associated with a concept shows a list of that concept along with other similar concepts that the user may want to examine.
- Users can check additional related concepts or substitute them for the original by unchecking it.

PDQ Interaction - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites History Print

Address <http://192.168.0.5/cgi-bin/PDQAgentX> Go

Continue with query refinement:

You can modify your original question:

You appear to be interested in the variable, sex. That variable and related variables and categories are listed here. Check all that are of interest. Be sure to include the variable sex if it should be included.

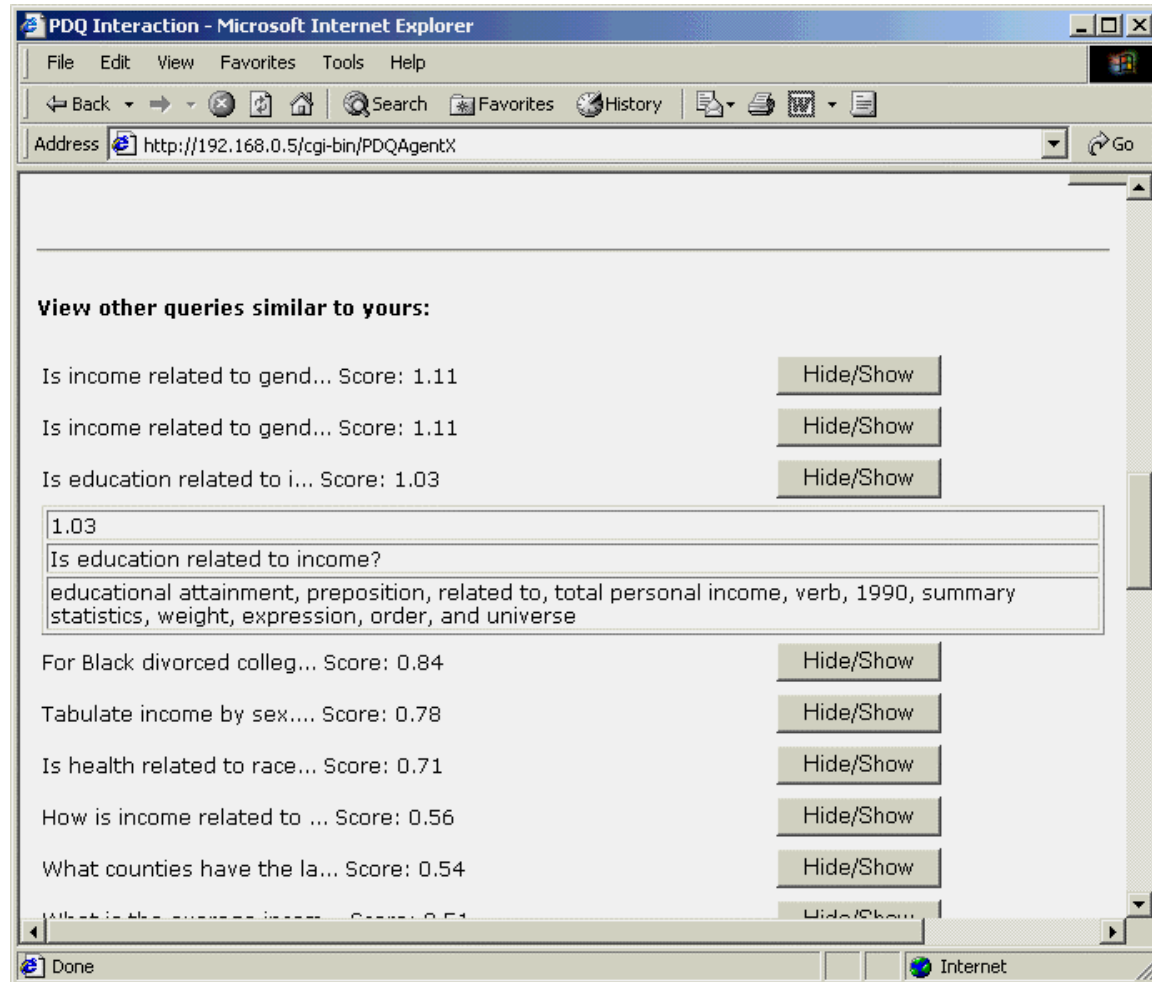
You appear to be interested in the variable, total personal income. That variable and related variables and categories are listed here. Check all that are of interest. Be sure to include the variable total personal income if it should be included.

<input checked="" type="checkbox"/> total personal income	<input type="checkbox"/> wage and salary income	<input type="checkbox"/> business and farm income
<input type="checkbox"/> non-farm business income	<input type="checkbox"/> farm income	<input type="checkbox"/> social security income
<input type="checkbox"/> welfare (public assistance) income	<input type="checkbox"/> interest, dividend, and rental income	<input type="checkbox"/> retirement income
<input type="checkbox"/> other income	<input type="checkbox"/> total personal earned income	<input type="checkbox"/> affluent
<input type="checkbox"/> poverty status	<input type="checkbox"/> middle-income	<input type="checkbox"/> income measure
<input type="checkbox"/> had nonwage/salary income over \$50	<input type="checkbox"/> deductions for retirement	<input type="checkbox"/> wealth measure
<input type="checkbox"/> real estate value	<input type="checkbox"/> value of personal estate	<input type="checkbox"/> well off
<input type="checkbox"/> employed	<input type="checkbox"/> educational attainment	<input type="checkbox"/> marital status

Done Internet

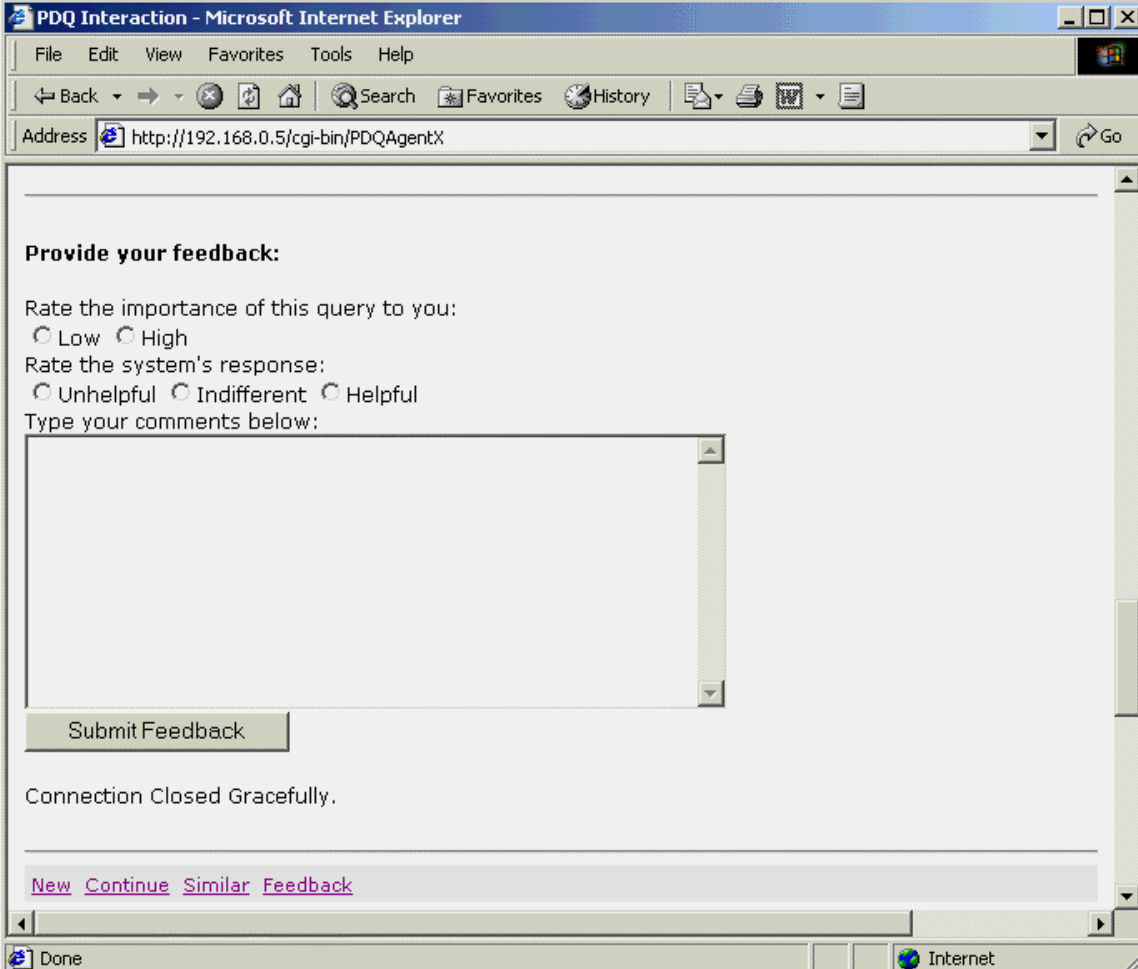
Similar Cases

- The next part of the feedback to the user shows a list of previous questions ordered with the ones most like the current displayed at the top of the list.
- Clicking the “Hide/Show” button displays details for that previous question.
- Users can incorporate elements of previous similar queries into the current one.



User Assessment and Saving of the Result in the Case File

- The final step in the CBR process is to assess the user's satisfaction with the result, then save the resulting query along with the original question and all of its relevant parameters as a case in the database.
- “Successful” cases will be given scores that encourage them to be displayed in the future, while “unsuccessful” cases will be used to help avoid repeating past mistakes.



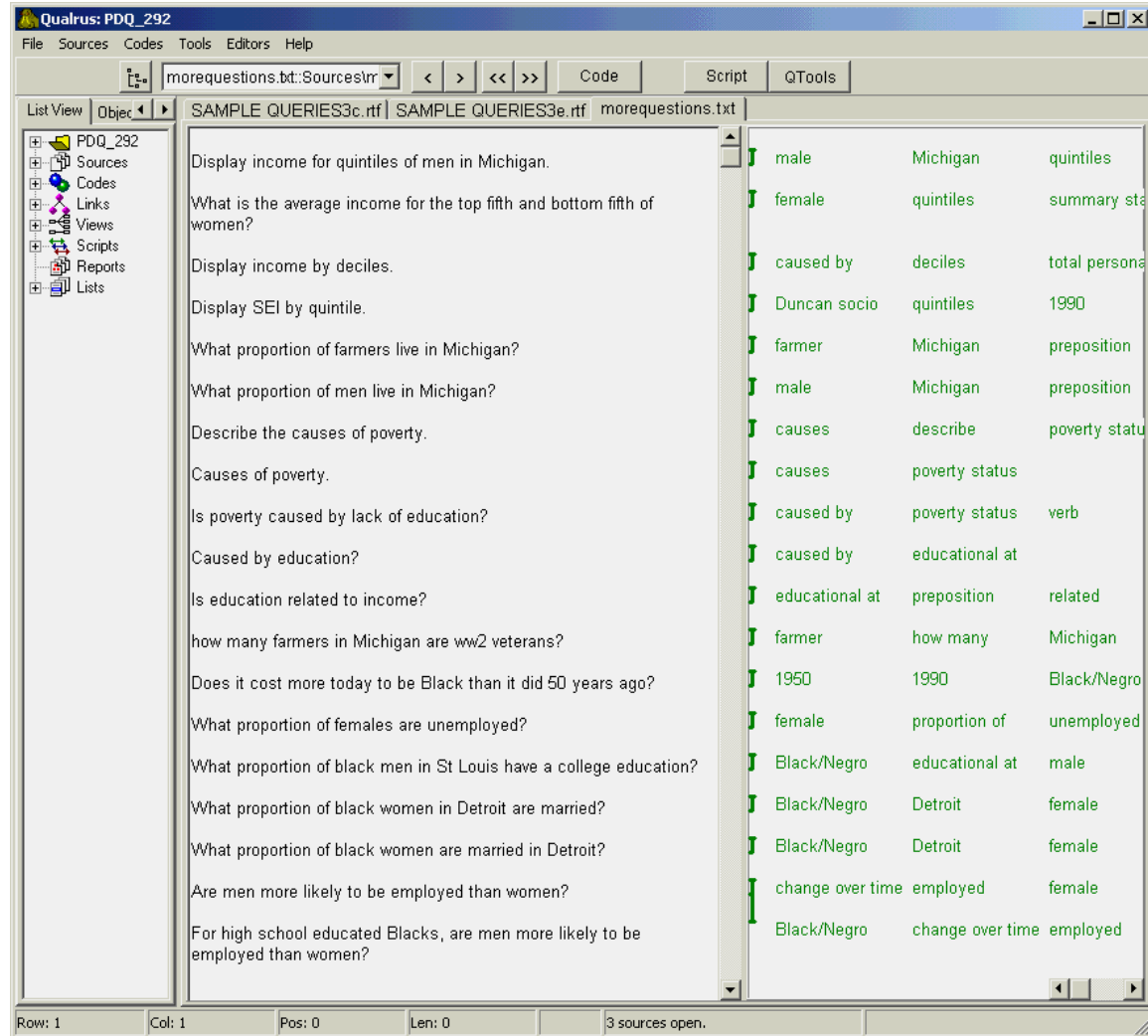
The screenshot shows a Microsoft Internet Explorer browser window titled "PDQ Interaction - Microsoft Internet Explorer". The address bar displays "http://192.168.0.5/cgi-bin/PDQAgentX". The main content area contains a feedback form with the following elements:

- Provide your feedback:**
- Rate the importance of this query to you:
 Low High
- Rate the system's response:
 Unhelpful Indifferent Helpful
- Type your comments below:
-

Below the form, the text "Connection Closed Gracefully." is displayed. At the bottom of the page, there are links: [New](#) [Continue](#) [Similar](#) [Feedback](#). The browser's status bar at the bottom shows "Done" and "Internet".

The Database of Previous Cases

- Each query, including both the text of the question and all codes associated with the question describing the resulting query, is saved in the database for consideration when the user enters future questions.



CBR vs. Other Learning Strategies (1)

- Neural networks, genetic algorithms, and other learning strategies generalize by inducing a general set of rules for the complete range of possible problems even before the target problem is known (*“eager generalization”*) - a very difficult task
- CBR generalizes only implicitly by finding similar cases and does so only after the target problem is known (*“lazy generalization”*) - a much easier task. (Golding, 2002).

CBR vs. Other Learning Strategies (2)

- As a consequence of their different approaches to generalization, other learning strategies often require hundreds or thousands of cases and extensive pre-processing before they can be useful.
- CBR can be used immediately, even with few cases.

CBR vs. Other Learning Strategies

(3)

- CBR thus tends to be a good approach for domains where there are many ways to generalize a case and where lazy generalization can make the task much more manageable.

CBR vs. Rule-Based Systems (1)

- Rule-based systems represent knowledge explicitly as a series of related “if ... then ...” production rules, requiring considerable understanding and specification of the details of the knowledge.
- In contrast, CBR systems represent knowledge implicitly as a series of examples with the greatest demand for specificity the identification of the indices for comparing cases.

CBR vs. Rule-Based Systems (2)

- Rule-based systems are often “brittle,” failing badly when applied outside their domain of expertise or to problems beyond the scope of their original design.
- CBR can handle the full range of questions diverse users pose of the data, including questions program developers are unlikely to anticipate.

CBR vs. Rule-Based Systems

(3)

- Rule-Based systems often have a “knowledge bottleneck” in which development is impeded by the slow pace of developing rules.
- CBR systems can be used immediately and grow in power and utility as new cases are added.

PDQ-Expert: CBR Implemented in Tandem with Rule-Based Procedures

- The CBR strategy described here is implemented in tandem with rule-based procedures used to identify the key features of cases and specify the query.
- It provides an effective way for users to step outside the boundaries of the recommended solution to consider reasonable alternatives.

Current Status

- This case-based reasoning interface is currently in operation on the web and we are trying it out with a wide range of questions.
- We will be expanding our base of users testing the system shortly and hope to make this an operating component of the PDQ-series of programs in the coming year.

Summary and Overview

- This approach to helping users overcome their lack of knowledge about metadata appears to be a fruitful strategy that promises to provide a versatile and powerful interface to census and similar microdata.
- The approach has three specific strengths:

(1) Clarifying the Unknown

- Users (not just novice users) often do not know precisely what they want to ask, which data sets are relevant and appropriate to their concerns, and the characteristics of the items in the data sets.
- Users can be helped and their thinking clarified by viewing examples of similar questions and the queries they generate.

(2) Serving Diverse Users

- Novice and experienced users tend to ask very different kinds of questions and to need different kinds of help.
- This CBR system provides an effective way to supply diverse users with informative feedback and suggestions.

(3) Quick Implementation along with Continuing Improvement

- This CBR strategy can be put into place relatively quickly.
- It provides a framework for continued improvement in the knowledge of the system as new cases are added.

Literature Cited

- Golding, Andrew R.. 2002. *Case-Based Reasoning*. Upedia.com. The Free Encyclopedia. <http://www.nupedia.com/article/short/Case-Based+Reasoning>
- Golding, A.R. and P.S. Rosenbloom. 1993. *Improving Rule Based Systems Through Case-Based Reasoning*. Pages 759-764 in Buchanan, B.G. and Wilkins, D.C. (Eds.) *Readings in Knowledge Acquisition and Learning: Automating the Construction and Improvement of Expert Systems*. Morgan Kaufmann Publishers, Inc., San Mateo, CA.
- Kolodner, Janet, L. and David B. Leake. 1996. In Leake, David B. (Ed.) *Case-Based Reasoning: Experiences, Lessons, and Future Directions*. Menlo Park, CA: American Association for Artificial Intelligence Press.

Thank You

- Edward Brent, Idea Works, Inc.
 - www.ideaworks.com
- Albert Anderson, Public Data Queries, Inc.
 - www.pdq.com
- Lisa Neidert, University of Michigan
 - www.psc.isr.umich.edu
- Pawel Slusarz, Idea Works, Inc.
 - www.ideaworks.com