From: AAAI Technical Report SS-96-05. Compilation copyright © 1996, AAAI (www.aaai.org). All rights reserved.

Improving FAQfinder's Performance: Setting Parameters by Genetic Programming

Edwin Cooper
University of Chicago AI Lab
1100 East 58th St.
Chicago, IL 60637
cooper@cs.uchicago.edu

Introduction: the FAQfinder System

As the quantity of information available on the Internet continues to increase, so do the attempts to harness that information. One such attempt is the University of Chicago's FAQfinder project; given a natural language English question, the FAQfinder system attempts first to locate the most relevant list of Frequently Asked Questions (FAQs), and then to find the question on that list most similar to the one asked by the user. Since FAQs number in the thousands and each one represents a concentration of information in a specific subject area, the FAQfinder system offers a means of accessing useful information in a straightforward manner.

The FAQfinder system is currently available as an internal Web site to the University of Chicago Computer Science community. At present, there are twenty FAQs available to provide answers to user's questions. Research proceeds in several directions: currently, a procedure is being developed to automatically separate questions from answers in FAQ files, the number of FAQs available to the user is being expanded, and FAQfinder's ability to accurately determine the similarity of two natural-language questions is being improved. My own research focuses on this final line of research; in this paper, I describe an ongoing effort to use a machine learning technique (in this case, John Koza's Genetic Programming) as a means of improving FAQfinder's question-matching performance.

FAQfinder's Functionality: the Process of Matching User Question to FAQ Question

After posing a natural language question to the FAQfinder system, the visitor to FAQfinder's web pages is asked to choose among the five FAQs most relevant to his or her question. The FAQfinder system then provides the user with the FAQ question that has been matched to the user question, along with its answer. This matching process is the combined result of three different strategies: the proximity of term-vectors, the comparison of pre-defined 'question-types', and the results from a marker-passing system operating over data extracted from Princeton's WordNet semantic network.

The selection of the five most relevant FAQs is accomplished using Cornell's SMART system. A term vector approach is also used within the second stage of lookup, as a means of matching the user's question to the most similar question within the most relevant FAQ. A standard measure is taken of the proximity of the user-question's term-vector to the term-vector of the FAQ question-answer pair; nearby vectors are taken as an indication that the two questions are also semantically similar.

A second factor in question matching is the use of a question grammar. Both the user's question and the FAQ questions are parsed into any of eight question-types (for example, one recently asked question, "How do I obtain my credit report?", was tagged by FAQfinder as a "how" question). The notion here is that similar questions are often phrased in similar ways; a "where" question is more likely to match another "where" question than to match a "how" or "what" question. If, at runtime, a definite question type is found for the user question, FAQ questions of identical question-type are considered closer matches to the user question.

A final indication of the match between two questions is obtained by passing markers over data obtained from the WordNet semantic network. Each entry ("synset") within the WordNet network is connected by a set of links to other, semantically related entries. One robust set of links within WordNet are the hypernym / hyponym (is-a) relations. In a system similar to Quillian's original model (Quillian 1972), activation spreads from words within the user question via is-a links supplied by WordNet to other, hierarchically related words. The shallow lexical semantics provided by the WordNet links allow semantically related words to add to questions' similarity; if two questions contain many words with the same part of speech and nearby positions in the WordNet network, they are considered to be more likely to match. One example of the successful application of this technique is for the questions, "How do I get my wife off my credit report?", and "How can I get my husband off my credit report?".

The words "husband" and "wife" are nearby in the WordNet network (both are connected by is-a links to the "spouse" intermediary); this proximity of words within WordNet adds to the closeness of match of the questions as a whole.

Optimizing FAQfinder: a Problem in Machine Learning

While these three strategies broadly define FAQfinder's ability to determine the similarity between any two questions, they say little about the relative weights which each strategy should receive. Each strategy incorporates a vector of controlling parameters; for instance, the question-matching strategy relies on a threshold of certainty in determining a question type, below which question-type is undefined. The marker passing strategy, on the other hand, incorporates an outermost limit for the spread of activation, and a rate of decay between nodes as activation spreads. While the limits of FAQfinder's performance are defined by this set of parameters involved in question-matching, its performance within that range is defined by the weighted combination of these parameters. A successful matching strategy within this search space will heavily weight a strong indication of question matching (such as term-vector proximity), weakly weight other indications, and perhaps take the ratio of a third and fourth parameter into account. FAQfinder's performance depends, effectively, on the nonlinear weighted combination of this set of question-matching parameters.

I address the task of locating an optimal weighted nonlinear combination of question-matching parameters as a problem in machine learning. As a machine learning technique, I have adopted John Koza's Genetic Programming (Koza 1994). This technique, similar in character to John Holland's Genetic Algorithms, allows the evolution of (possibly nonlinear) functions, rather than the simple adjustment of a vector of weights. Mutation and crossover occur among the population of functions, and those functions which define superior weighted combinations of parameters are given a relatively greater chance of reproducing. In anticipation of this research, the current FAQfinder system has been adapted to this purpose; as one input, it takes a Lisp function defining how each of the parameters involved in question matching should be weighted and combined to determine the proximity between two questions.

As a machine learning problem, the task of combining question-matching parameters is unusually well defined. We have available a suite of over one-hundred sample user questions, each of which was tagged by hand to indicate which FAQ question(s) it should match. Given a Lisp function which defines a weighted combination of question-matching parameters, we therefore have an immediate indication of FAQfinder's resulting performance. In its current version, at its highest level of recall, FAQfinder is correctly matching 80% of the test questions. By using an initial population containing the function which defines the current matching strategy, the FAQfinder system will at least maintain this level of performance, and hopefully exceed it.

It should be noted that this project is oriented toward the empirical improvement of FAQfinder's performance. It will not be sufficient simply to improve slightly on the current weighted combination of parameters. The project will be considered successful if and only if a nonlinear combination of FAQfinder parameters is found which produces a better performance than could be reasonably expected from a by-hand weighted combination of question-matching parameters. This is an emphasis on machine learning for empirical improvement, rather than machine learning for its own sake.

The AAAI Spring Symposium poster session will include a demonstration of FAQfinder, whose weighted combination of parameters (and resulting performance) is obtained by Genetic Programming. Of special interest may be the application of machine learning to marker-passing; the success of this technique will be illustrated and discussed.

References

Collins, A. M. and Quillian, M. R. (1972). How to Make a Language User. In E. Tulving and W. Donaldson, Organization of Memory. New York: Academic Press.

Koza, John R. 1994a. Genetic Programming 11: Automatic Discovery of Reusable Programs. Cambridge, MA: The MIT Press.

Previous work on FAQfinder:

Hammond, K.J.; Burke, R., and Schmitt, K. (1994). A Case-Based Approach to Knowledge Navigation. In AAAI Workshop on Knowledge Discovery in Databases. AAAI. August 1994. Seattle WA.

Hammond, K.J.; Burke, R., Martin, C., and Lytinen, Stephen (1995). FAQ Finder: A Case-Based Approach to Knowledge Navigation. In AAAI Symposium on Information Gathering in Heterogeneous, Distributed Environments. AAAI. March 1995. Stanford University.