Machine Discovery Based on Numerical Data Generated in Computer Experiments

Tsuyoshi Murata and Masamichi Shimura
Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology
2-12-1 Oh-okayama, Meguro, Tokyo 152, JAPAN
{murata, shimura}@cs.titech.ac.jp

Abstract

In the discovery of useful theorems or formulas, experimental data acquisition plays a fundamental role. Most of the previous discovery systems which have the abilities for experimentation, however, require much knowledge for evaluating experimental results, or require plans of common experiments which are given to the systems in advance. Only few systems have been attempted to make experiments which enable the discovery based on acquired experimental data without depending on given initial knowledge.

This paper proposes a new approach for discovering useful theorems in the domain of plane geometry by employing experimentation. In this domain, drawing a figure and observing it correspond to making experimentation since these two processes are preparations for acquiring geometrical data. EXPEDITION, a discovery system based on experimental data acquisition, generates figures by itself and acquires expressions describing relations among line segments and angles in the figures. Such expressions can be extracted from the numerical data obtained in the computer experiments. By using simple heuristics for drawing and observing figures, the system succeeds in discovering many new useful theorems and formulas as well as rediscovering well-known theorems, such as power theorems and Thales' theorem.

Introduction

Since the beginning of human history, scientists have discovered many useful theorems and formulas from the data acquired by experimentation. Zytkow and Baker (Zytkow & Baker 1991) pointed out the advantages of experimentation for discovery: 1) experimentation provides an abundance of data, 2) extremely accurate data can be acquired by experimentation, 3) an experimenter can create special situations that are otherwise not available, and 4) an experimenter can create simple experimental situations so that empirical regularities are easy to discover. It is expected that the abilities of experimentation carry the above advantages to a discovery system as well as to a human scientist.

Although many discovery systems focus on experimentation as a method of interaction with the external world, most of the systems require considerable amount of given knowledge. KEKADA (Kulkarni & Simon 1988) focuses its attention on surprising phenomena to constrain the search space of experimentation. In order to detect surprising phenomena, however, the system needs to have knowledge about ordinary experimental results. DEED (Rajamoney 1993) designs experiments which discriminate between two competing theories. Since the system is based on the difference of causal explanations by the competing theories, it is not applicable to the situations where there exist no such theories.

In order to make experiments, the abilities of planning experimental procedures are important. Such abilities have been incorporated in some discovery systems, including MOLGEN (Friedland 1979) and STERN (Cheng 1992), which employ experimentation. Most of these systems, however, need to have prescribed domain-dependent experimental plans which are given to the systems in advance.

For discovering new theorems by a deductive process, domain knowledge is very important to generate a set of theorematic candidates. Furthermore many heuristics are needed to plan the experiment for obtaining appropriate data and to avoid computational explosion in a search space. In the knowledge-intensive systems such as AM (Lenat 1983), one of the well-known discovery systems, given knowledge is combined or mutated to generate new theorems for discovery. Since the generation of desired theorems heavily depends on the given knowledge, there is the possibility that the system may not be able to discover useful theorems according to the lack of given knowledge. Also the theorems and formulas discovered are sometimes restricted in domain, since the possible methods of experimentation depend entirely on such knowledge.

As described before, experimentation generally provides an abundance of data from external environment. Discovery based on experimental data acquisition, therefore, is more desirable for a discovery system. This is because the method is expected to make up for missing initial knowledge by discovering knowledge from observed experimental data. Especially in the domain of plane geometry, much data from which useful theorems are extracted can be obtained by drawing figures and by finding their geometrical relations.

As Shrager and Langley (Shrager & Langley 1990) pointed out, a discovery system for mathematics is unusual compared with a system for physics or chemistry, in that the system can generate data internally rather than observing them in a real or simulated environment. This means that a discovery system for mathematics has the property of making internal experiments with less knowledge of experimentation than the systems for other domains. A discovery system for plane geometry, which is our target, is also able to take the advantage of the property by generating figures and observing them in the system.

This paper proposes a new approach for discovering useful theorems in the domain of plane geometry by employing experimentation. EXPEDITION, a discovery system based on EXPErimental Data acquisITION, generates figures automatically by drawing lines one by one, and observes the figures in order to extract numerical data. From the numerical data, expressions about line segments and angles are acquired. Although many expressions are acquired from a figure, the expressions about line segments and angles which are newly generated by the last additional line are regarded as useful in the system since the expressions cannot be acquired from the figure before drawing the line. With only two simple heuristics for drawing and observing figures, EXPEDITION succeeds in discovering many useful theorems as well as rediscovering well-known theorems such as power theorems and Thales' theorem.

Discovery based on the comparison of experimental results

In order to clarify the role of experiments for discovering knowledge, the processes of actual discovery have been investigated. The records of actual discovery processes, such as laboratory notes and recollections of a discoverer, have been often used as the bases for developing discovery systems. There are two approaches to the study of actual human discovery. One involves the analysis of historical records of real scientists, and the other involves the analysis of the behavior of subjects who are working on a discovery task, such as a task of discovering the mechanism of a device or a chemical reaction.

Dunbar (Dunbar 1993) analyzed the experimental processes of subjects who were asked to discover how genes were controlled by using a simulator of genes. Klahr et al. (Klahr, Dunbar, & Fay 1990) used a computer-controlled robot tank, which can be programmed with a sequence of commands, as a device for a discovery task. Subjects were asked to discover the operation of an unknown command. When they observed the behavior of the robot tank whose program included the unknown command, most of them realized that a part of the commands in the program was executed repeatedly. Then they executed similar programs whose numerical parameters were different from the previous program, and compared the results to clarify the range of the repetition. This conservative strategy is called the VOTAT (vary one thing at a time) experimental strategy. Schunn and Klahr (Schunn & Klahr 1995) obtained experimental data using a simulator called MilkTruck. Subjects of this research also conducted a sequence of similar programs for the discovery of the operation of unknown commands.

As is seen above, analyzing and comparing each result obtained by similar experiments are very important in evaluating the results and in discovering new knowledge or theorems, even if initial background knowledge is not fully available. Such a mechanism of comparison, therefore, is essential and desirable also in a discovery system employing experimentation in order to detect regularity and peculiarity of the results obtained. In the domain of plane geometry, therefore, a system which operates based on comparison of its experimental results is expected to be able to discover useful desired theorems from various data of figures generated in the experiments.

Discovery based on experimental data acquisition

Drawing figures

Figures often enable the detection of visual information such as neighborhood relations and relative size. This property is called emergent property (Koedinger 1992), which is one of the reasons humans use figures for solving problems. By drawing figures and observing them, a discovery system for plane geometry is also able to acquire much geometrical data.

In order to draw various figures for the acquisition of data, lines are added one by one on a given base figure. In this paper, a circle is chosen as the base figure since many interesting figures can be drawn from a circle. To guide line drawing, *focus points* are introduced such as the center of a circle, a point on the circumference, contact points, and intersection points. Lines are drawn in the following way according to the focus points:

From a focus point outside the circle

- draw a tangential line to the circle
- draw a line which passes through the center of the circle
- draw an arbitrary line which has common points with the circle

From a focus point on the circle circumference

- draw a tangential line which touches the circle at the focus point
- draw a line which passes through the center of the circle
- draw an arbitrary line to a point on the circumference

From a focus point inside the circle

- draw a line which passes through the center of the circle
- draw an arbitrary line which has common points with the circle

Figure 1 shows a part of drawn figures in the above way. Dots in the figures indicate focus points.

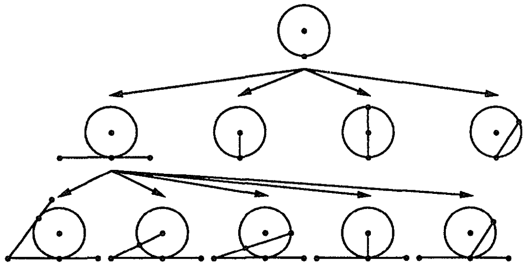

Figure 1: Drawing figures by adding lines on a circle

Acquisition of theorematic candidates

By observing the figure drawn in the above procedure, numerical data are acquired such as the length of line segments and the measure of angles. The length of line segments, and the sum and the product of the length of two arbitrary line segments are listed from the data. An expression, which we call a theorematic candidate, is acquired from two entries of approximately equal numerical values in the list. For example, in the figure shown in Figure 2, a theorematic candidate $AB^2 = AD \cdot AE$ is acquired based on the observed numerical data.

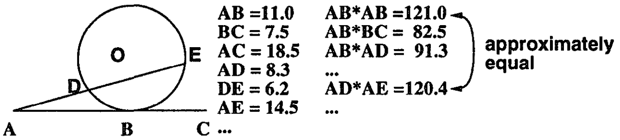

Figure 2: Acquisition of a theorematic candidate

From the data of angles also, theorematic candidates are acquired in the same manner. The following obvious relations are included in the acquired theorematic candidates, which means our approach succeeds in discovering the relations.

- Radii (diameters) of a circle are equal.
- A diameter is twice a radius.
- The sum of divided lines is equal to the original line.
- If A, B and C are three collinear points, the measure of an angle $\angle ABC$ is 180°.
- The sum of divided angles is equal to the original angle.
- The sum of the measures of three angles in a triangle is 180°.

Selection of useful theorematic candidates

Many theorematic candidates are acquired from a figure. As additional lines are drawn on a figure, the number of line segments and angles increases, and then the combination of line segments and angles increases accordingly. As a result, numerous theorematic candidates can be acquired from a complicated figure composed of many lines. To obtain only useful theorems from many acquired theorematic candidates, it is important to select useful theorematic candidates.

Let us focus on the relations about line segments and angles which are newly generated by drawing an additional line on a figure. Since theorematic candidates about the newly generated line segments and angles cannot be acquired from the figure before drawing the additional line, such candidates can be considered as useful.

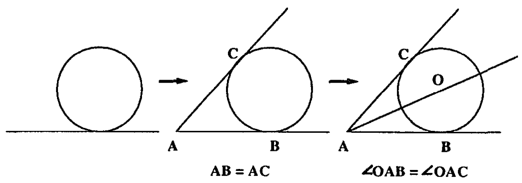

Figure 3: Selection of useful theorematic candidates

Figure 3 shows a sequence of figures and corresponding useful theorematic candidates. In the middle figure of Figure 3, an expression $AB = AC$ is regarded as a useful theorematic candidate since it shows the relation about newly generated line segment AC. In the right figure, an expression $\angle OAB = \angle OAC$ is regarded as useful for the same reason. By focusing on the relations about newly generated line segments and angles, the combinatorial explosion is avoided and the discovery from a complicated figure is also enabled.

DST (Murata, Mizutani, & Shimura 1994) is one of the discovery systems in the domain of plane geometry. It discriminates line segments and angles which are generated by auxiliary lines. Such line segments and angles, called subproducts, are eliminated from acquired expressions by transformation to discover theorems which include no subproduct. DST draws auxiliary lines only for the purpose of extracting the data of line segments and angles which already exist before drawing the lines. On the other hand, the approach proposed here draws additional lines for the purpose of extracting the data of newly generated line segments and angles. Since additional lines are regarded as constituents in a figure, our new approach enables the discovery from various figures.

Verification of theorematic candidates

The theorematic candidates which hold only for the original figure, the figure from which they are acquired, are not true theorems. To remove such candidates, every candidate from the original figure should be tested whether the candidate holds for other figures which topologically resemble the original figure. Such figures are re-drawn by adding lines in the same order as the original figure. This is because the figures are used for making other experiments which resemble the one using the original figure. Since an additional line is drawn at random in length and in direction, re-drawn figures are, in general, partly different from the original figure. As a result of the above experiments, a theorematic candidate which holds for all the figures is regarded as a useful theorem of great generality.

Repetitive experiments are often carried out by human scientists as well in order to test whether an observed surprising phenomenon of a substance is exhibited generally by other substances of the same class. Such experiments are necessary for assessing the scope of the phenomenon. Drawing the figures which resemble the original figure can be considered as making supplementary experiments for verifying the generality of discovered theorems. However, unlike the repetitive experiments of previous discovery systems, re-drawing figures is very simple and requires less domain knowledge.

Experimental results

We have developed EXPEDITION, a discovery system based on experimental data acquisition, by using the proposed approach mentioned above. EXPEDITION succeeds in discovering many useful theorems as well as rediscovering well-known theorems about the figures which include a circle. Figure 4 shows some of the figures generated in our system. From these figures, the following well-known theorems are rediscovered by interpreting acquired expressions:

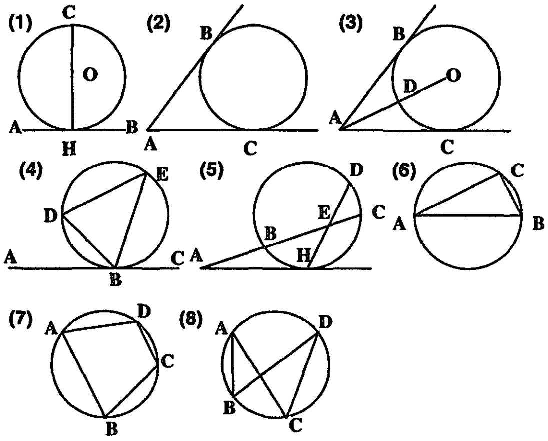

Figure 4: Figures for rediscovering theorems

1. A tangential line to a circle is perpendicular to the radius (diameter) from the contact point. ($\angle AHO = 90°$)
2. Two line segments from a point outside a circle to its contact points are equal. ($AB = AC$)
3. A line from the vertex of an angle to the center of inscribed circle is a bisector of the angle. ($\angle OAB = \angle OAC$)
4. An angle of the triangle inscribed in a circle is equal to an angle between the chord opposite to the angle and the tangential line which touches the circle at the end point of the chord. ($\angle DEB = \angle DBA, \angle EDB = \angle EBC$)
5. Power theorems. ($AB \cdot AC = AH^2, BE \cdot EC = DE \cdot EH$)
6. Thales' theorem. ($\angle ACB = 90°$)
7. The sum of the measure of two opposite angles of an inscribed quadrilateral is 180°. ($\angle ABC + \angle CDA = 180^o$, $\angle BCD + \angle DAB = 180^o$)
8. Inscribed angles in a circle are equal when their end points of sides excluding their vertices are the same. ($\angle BAC = \angle BDC, \angle ABD = \angle ACD$)

Moreover, EXPEDITION discovers many other theorems which are not found in a conventional book of geometry. From the figures shown in Figures 5 and 6, the following theorems (1) and (2) are discovered respectively:

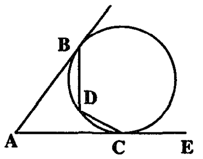

Figure 5: A figure for discovering theorem (1)

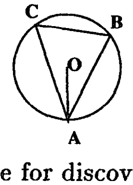

Figure 6: A figure for discovering theorem (2)

$$\angle ABD + \angle BDC = \angle DCE \quad (1)$$
$$\angle OAC + \angle ABC = 90° \quad (2)$$

Although these theorems can be proved easily, it is quite interesting that EXPEDITION draws the figures by itself and finds these expressions as useful ones. Many theorems about line segments are also discovered

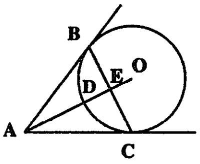

Figure 7: A figure for discovering theorem (3)

by the system. From the figure shown in Figure 7, the following simple and elegant theorem is discovered:

$$AD \cdot BE = AB \cdot DE \tag{3}$$

In order to deduce this theorem by using the geometrical relations such as similarity and congruence, addition of auxiliary lines and complicated transformation of expressions are required. The fact that EXPEDITION discovers such a theorem only from observed data shows that our approach is quite useful and that the system has advanced abilities for discovery.

From the figures shown in Figures 8 and 9, the following theorems { (4), (5) } and { (6), (7), (8), (9), (10), (11), (12) } are discovered respectively:

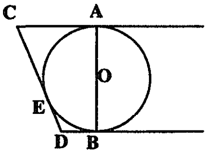

Figure 8: A figure for discovering theorems (4) and (5)

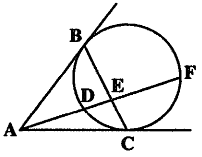

Figure 9: A figure for discovering theorems from (6) to (12)

$$\begin{aligned}
CD \cdot ED &= OB^2 + ED^2 && (4)\\
CE \cdot ED &= OB^2 && (5)\\
AE \cdot AF &= AC^2 + AD \cdot EF && (6)\\
AD \cdot AF &= AE^2 + CE \cdot EB && (7)\\
AD \cdot AF &= AE^2 + DE \cdot EF && (8)\\
EF \cdot AF &= AD \cdot DE + DF^2 && (9)\\
AE \cdot EF &= DE \cdot AF + CE \cdot EB && (10)\\
AD \cdot EF &= AE \cdot DE + CE \cdot EB && (11)\\
AD \cdot EF &= DE \cdot AF && (12)
\end{aligned}$$

It must be noted that the discovery from the figures which include no similar or congruent triangles, such as Figures 8 and 9, is also realized in our system.

Discussion

Most of the previous discovery systems employ heuristics for controlling their search in order to avoid the combinatorial explosion. Such heuristics, however, often require considerable amount of knowledge. In EXPEDITION, only the following two heuristics are used:

- drawing figures by adding lines
- focusing on the expressions about line segments and angles which are newly generated by the last additional line

The former enables the system to draw various figures automatically and to acquire data by observing the figures. The latter avoids the combinatorial explosion without using knowledge for search. Although the above both heuristics of EXPEDITION do not require domain knowledge, they both contribute very much to the discovery of many useful theorems.

Figures are generated from simple ones to complicated ones by drawing lines one by one. Generating figures in this way enables the system to discover theorems about various figures efficiently without using much knowledge. The system sets up hypotheses about the relations among line segments and angles, what we call theorematic candidates in this paper, by using numerical data acquired from a figure. In order to verify the theorematic candidates, many figures which resemble the original figure are re-drawn in the same order as an original figure. Since the above experiments are made internally, the system does not need to have much knowledge for experimentation. This approach is based on a generate-and-test procedure and is suitable for machine discovery.

In general, geometrical theorems are often deduced using the expressions acquired from the geometrical relations such as similarity and congruence. In order to deduce theorems from a figure which has no such geometrical relations, auxiliary lines which generate the relations have to be drawn on the figure. However, drawing appropriate auxiliary lines is a very difficult and tricky task. By using numerical data, EXPEDITION discovers theorems which are difficult to deduce only from the expressions of the geometrical relations.

In the domains of physics and chemistry, an expression which holds for all similar experiments is considered as a true law. Similarly, an expression which holds for all re-drawn figures is regarded as a true theorem in our system. Practically, there is no need to re-draw figures many times; only a few times of re-drawing are enough for verifying theorematic candidates. From the figure shown in Figure 9, for example, 23 nontrivial theorematic candidates were hypothesized at first. By observing only one re-drawn figure, 10 candidates were invalidated and all the remaining ones, including the theorems described before, were actually true.

In order to deduce geometrical theorems, some of the previous work on theorem proving, such as Gelernter's geometry-theorem proving machine (Gelernter 1963) and DC model (Koedinger and Anderson 1990), also use geometrical data which are observed from figures. Gelernter's system uses figures to prune invalid geometrical relations that are generated by the backward search. In the DC model, figures are used to generate hypotheses which are pruned by using domain knowledge. Our approach is different from the above both approaches in that figures are used for both generating hypotheses and validating them. Therefore, EXPEDITION is able to acquire theorems without depending on given domain knowledge.

Conclusion

We have described an approach for discovering useful theorems in the domain of plane geometry by employing experimentation. EXPEDITION, which we developed, succeeds in discovering many useful theorems as well as rediscovering well-known theorems such as power theorems and Thales' theorem.

The success of EXPEDITION shows that experimentation plays an important role in discovering theorems. In general, an empirical method of scientific discovery requires several processes such as making experimental plans, acquiring data by experimentation, setting up appropriate hypotheses, and verifying the hypotheses. Since our system draws figures, which corresponds to making experimentation, it does not need to have knowledge for making experimental plans; it discovers theorems by using nothing but the heuristics of drawing figures and the heuristics of focusing on newly generated line segments and angles.

In the domains of physics and chemistry, numerous experimental data which are acquired based on domain knowledge are used for discovering useful laws. A discovery system which simulates human discovery processes in such domain requires much knowledge and heuristics. On the other hand, in the domain of mathematics, especially in plain geometry, expressions acquired from domain axioms or from observed figures are used with insight for discovering theorems and formulas. In other words, laws in physics and chemistry are discovered inductively while theorems in mathematics are discovered deductively. Although EXPEDITION acquires expressions from numerical data rather in an inductive way, the system actually discovers novel theorems in the domain of plane geometry without using much knowledge. Such inductive discovery is desirable for various domains in which computer-controlled experimentation is available.

References

Cheng, P. C.-H. 1992. Diagrammatic Reasoning in Scientific Discovery: Modelling Galileo's Kinematic Diagrams. Technical Report SS-92-02, 1992 AAAI Spring Symposium, Reasoning with Diagrammatic Representations, 33 - 38.

Dunbar, K. 1993. Concept Discovery in a Scientific Domain. *Cognitive Science* 17(3):397 – 434.

Friedland, P. 1979. Knowledge-based Experiment Design in Molecular Genetics. In *Sixth International Joint Conference on Artificial Intelligence*, 285 – 287.

Gelernter, H. 1963. Realization of a geometry-theorem proving machine. In Feigenbaum, E. A., and Feldman, J., eds., *Computers and Thought.* McGraw-Hill. 134 - 152.

Klahr, D.; Dunbar, K.; and Fay, A. L. 1990. Designing Good Experiments To Test Bad Hypotheses. In Shrager, J., and Langley, P., eds., *Computational Models of Scientific Discovery and Theory Formation.* Morgan Kaufmann. chapter 12, 355 – 402.

Koedinger, K. R., and Anderson, J. R. 1990. Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science* 14(4):511 – 550.

Koedinger, K. R. 1992. Emergent Properties and Structural Constraints: Advantages of Diagrammatic Representations for Reasoning and Learning. Technical Report SS-92-02, 1992 AAAI Spring Symposium, Reasoning with Diagrammatic Representations, 151 - 156.

Kulkarni, D., and Simon, H. A. 1988. The Processes of Scientific Discovery: The Strategy of Experimentation. *Cognitive Science* 12(2):139 – 175.

Lenat, D. B. 1983. The Role of Heuristics in Learning by Discovery: Three Case Studies. In Michalski, R. S.; Carbonell, J. G.; and Mitchell, T. M., eds., *Machine Learning : An Artificial Intelligence Approach.* Tioga. 243 – 306.

Murata, T.; Mizutani, M.; and Shimura, M. 1994. A Discovery System for Trigonometric Functions. In *Proceedings, Twelfth National Conference on Artificial Intelligence*, 645 – 650. The AAAI Press.

Rajamoney, S. A. 1993. The Design of Discrimination Experiments. *Machine Learning* 12:185 – 203.

Schunn, C., and Klahr, D. 1995. A 4-Space Model of Scientific Discovery. Technical Report SS-95-03, 1995 AAAI Spring Symposium, Systematic Methods of Scientific Discovery, 40 - 45.

Shrager, J., and Langley, P. 1990. Computational Approaches To Scientific Discovery. In Shrager, J., and Langley, P., eds., *Computational Models of Scientific Discovery and Theory Formation.* Morgan Kaufmann. chapter 1, 1 – 25.

Zytkow, J. M., and Baker, J. 1991. Interactive Mining of Regularities in Database. In Piatetsky-Shapiro, G., and Frawley, W. J., eds., *Knowledge Discovery in Databases.* The AAAI Press. 31 – 53.