Data Mining the
Online Encyclopedia of Integer Sequences
for New Identities

Hieu Nguyen

Rowan University

MAA-NJ Section Spring Meeting
March 31, 2012

Acknowledgements

Doug Taggart (Undergraduate Research Assistant)

OEIS Deluge

- **Online Encyclopedia of Integer Sequences (OEIS)**

1. Searchable online database - http://oeis.org
2. Contains over 200,000 integer sequences
3. Created by Neil Sloane (AT & T Bell Labs), currently maintained by OEIS Foundation
4. Example: $F_n = 0, 1, 1, 2, 3, 5, 8, 13, 21, \ldots$

Mining the OEIS

- **Data Mining (Large Scale Pattern Recognition)**

  Process of extracting patterns from large data sets using computer science, mathematics, and statistics.

- **Mine OEIS for Integer Sequence Identities**

  1. Enlarge OEIS database to include transformations of integer sequences
  2. Find matches between sequence transformations (experimental conjectures)
  3. Prove experimental conjectures that are interesting to obtain new identities
  4. GOAL: Discover interesting connections between different areas of mathematics

Experimental Pattern Matching

- **Example 1**
- **A000045: Fibonacci sequence; $F(n) = F(n-1) + F(n-2)$, $F(0) = 0$, $F(1) = 1$**
  $F(n) = 0, 1, 1, 2, 3, 5, 8, 13, 21, \ldots, 39088169$ (39 terms); $n \geq 0$

1. A000045S1T3: Sums of Squares Transformation
   $G(n) = \sum_{k=0}^{n} F(n)^2 = 0, 1, 2, 6, 15, 40, 104, \ldots, 2472169789339634;\ n \geq 0$

2. A000045S1T8: Product of Consecutive Terms Transformation
   $H(n) = F(n) \cdot F(n+1) = 0, 1, 2, 6, 15, 40, 104, \ldots, 2472169789339634;\ n \geq 0$

3. Identical Match: $G(n) = H(n)$

EXPERIMENTAL CONJECTURE: $\boxed{\sum_{k=0}^{n} F_k^2 = F_n \cdot F_{n+1}}$

Example 2

- **A000295: Eulerian numbers (number of permutations of {1,2,...,n} with exactly one descent).**
  $a(n) = 0, 0, 1, 4, 11, 26, 57, 120, 247, 502, \ldots, 8589934558$; $n \geq 0$ (34 terms)

  1. A000295S1T9: Cassini Transformation:

  $G(n) = a(n+1)\,a(n-1) - a\,(n)^2 = 0, \text{-}1, \text{-}5, \text{-}17, \ldots, \text{-}3489660929$

- **A031878: Maximal number of edges in Hamiltonian path in complete graph on n nodes.**
  $b(n) = 0, 1, 3, 5, 10, 13, 21, 25, 36, \ldots, 1378$ $n \geq 1$ (53 terms)

  2. A031878S1T4: Binomial Transform of $b(n)$ (pad $b(0) = 0$):

  $H(n) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} b(k) = 0, 0, 1, 0, \text{-}1, \text{-}5, \text{-}17, \ldots, \text{-}3489660929, \ldots, \text{-}55169095435288577$

  3. Partial Match: $G(n) \approx H(n+3)$

  EXPERIMENTAL CONJECTURE:

  $$\boxed{a(n)^2 - a(n+1)\,a(n-1) = -\sum_{k=0}^{n+2} (-1)^{n+2} \binom{n+2}{k} b_k = (n-1)\,2^n + 1} \quad (n \geq 1)$$

Hunting for Identities

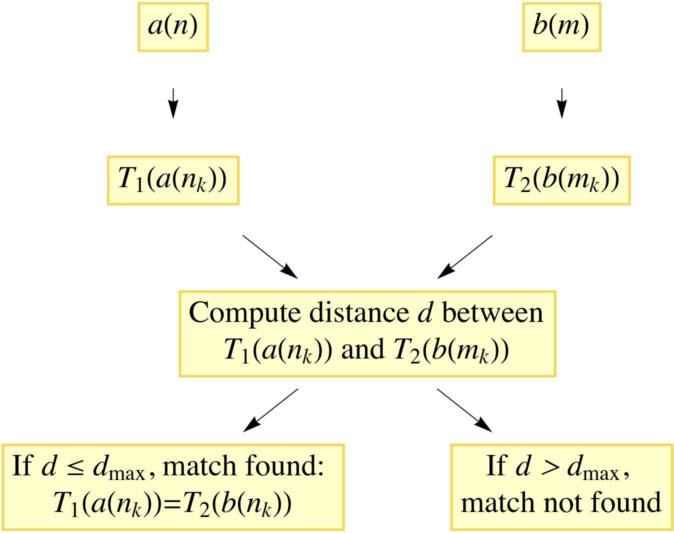
- **Classical Approach**

  Tools: Paper and pencil, good book-keeping
  Great bookkeepers: John Wallis, Isaac Newton, Leonard Euler

- **Modern Approach**

  Tools: Computers, computer algebra systems (e.g. Maple, *Mathematica*, Matlab, Sage)
  Small-scale: Search for identities one at a time using OEIS
  Large-scale: Mine for clusters of identities (EUREKA)

Pattern Matching Algorithm for Integer Sequences



$a(n)$

$b(m)$

$T_1(a(n_k))$

$T_2(b(m_k))$

Compute distance d between $T_1(a(n_k))$ and $T_2(b(m_k))$

If $d \leq d_{\max}$, match found: $T_1(a(n_k))=T_2(b(n_k))$

If $d > d_{\max}$, match not found

Database of Sequence Transformations

- **Raw Source Data - Sequences $\{a_n\}$ from OEIS**
- **Set of Transformations**

LABEL	TRANSFORMATION	FORMULA
T1	Identity	$a(n)$
T2	Partial Sums	$\sum_{k=0}^{n} a(k)$
T3	Partial Sums of Squares	$\sum_{k=0}^{n} a(k)^2$
T4	Binomial Transform	$\sum_{k=0}^{n} (-1)^k \binom{n}{k} a(k)$
T5	Self – Convolution	$\sum_{k=0}^{n} a(k)\, a(n-k)$
T6	Linear Weighted Partial Sums	$\sum_{k=1}^{n} k\, a(k)$
T7	Binomial Weighted Partial Sums	$\sum_{k=0}^{n} \binom{n}{k} a(k)$
T8	Product of Consecutive Elements	$a(n)\, a(n+1)$
T9	Cassini	$a(n-1)\, a(n+1) - a(n)^2$
T10	First Stirling	$\sum_{k=0}^{n} s(n, k)\, a(k)$
T11	Second Stirling	$\sum_{k=0}^{n} S(n, k)\, a(k)$

- **Create MySQL Database of Sequence Transformations**

ID	Label	Subsequence	Transformation	Position	Entry1	Entry2	Entry3
1	A000045S1T1	1	1	0	0	1	1
2	A000045S1T1	1	1	1	1	1	2
3	A000045S1T1	1	1	2	1	2	3
4	A000045S1T1	1	1	3	2	3	5
...	...	...	...	...	...	...	...
38	A000045S1T1	1	1	37	24 157 817	39 088 169	Null
39	A000045S1T1	1	1	38	39 088 169	Null	Null

1. Contains over 77 million rows (each row stores a window of 3 terms of a sequence) - 5 GB file

2. Contains extremely large numbers (up to 100 digits long)

3. Indexed to perform fast searches

Matching Integer Sequences

- **Main Assumption:**

  Perfect data set - no errors in the terms of each integer sequence

- **Challenges**

  1. Sequences vary in length (4 to 100 terms)

  2. High proportion of sequences begin with 0’s and 1’s.

  3. Find an effective similarity measure (i.e. distance function) to minimize ‘false matches’.

- **Overlapping Run**

  1. $\{\mathbf{1, 1, 2, 3, 5, 8, 13, 21}, 47, \mathbf{55}\}$
  $\{a(n)\} = \{\mathbf{1, 1, 2, 3, 5, 8, 13, 21}, 34, \mathbf{55}\}$
  NO MATCH (Worst)

  2. $\{\mathbf{55}, 89, 144, 233, 377, 610\}$
  $\{a(n)\} = \{1, 1, 2, 3, 5, 8, 13, 21, 34, \mathbf{55}\}$
  MATCH

  3. $\{\mathbf{3, 5, 8, 13, 21, 34, 55}, 89, 144, 233, 377\}$
  $\{a(n)\} = \{1, 1, 2, \mathbf{3, 5, 8, 13, 21, 34, 55}\}$
  MATCH

  4. $\{\mathbf{2, 3, 5, 8, 13, 21, 34}, 55\}$
  $\{a(n)\} = \{1, 1, \mathbf{2, 3, 5, 8, 13, 21, 34}\}$
  MATCH (Best?)

Head Bites Tail Overlap

- **What qualifies as a match between two finite sequences?**

$$\Big\{\overset{\text{Head}}{a(1)},\ a(2),\ \ldots,\ a(N-1),\ \overset{\text{Tail}}{a(N)}\Big\}$$

$$\Big\{\underset{\text{Head}}{b(1)},\ b(2),\ \ldots,\ b(M-1),\ \underset{\text{Tail}}{b(M)}\Big\}$$

We will say that two sequences *likely match* or are *similar* (in the sense that there is a chance that both finite sequences are part of the same infinite sequence) if the **head** (beginning) of one sequence **bites** (overlaps with) the **tail** (end) of the other sequence.

- **Head-Bites-Tails Overlap**

We say that two finite sequences contain a *head-bites-tail (HBT) overlap* if there is an overlapping run which starts at the beginning of one sequence and stops at the end of either sequence.

CASE 1:

$$a(1),a(2),\ldots\boxed{a(n_0),\ldots,a(N)}$$

$$\boxed{b(1),\ldots,\ b(L)},\ldots b(M)$$

CASE 2:

$$a(1),a(2),\ldots\boxed{a(n_0),\ldots,a(n_0+M-1)},\ldots,a(N)$$

$$\boxed{b(1),\ldots,\ b(M)}$$

HBT Distance

- **DEFINITION**

We define $L_{\max}$ to be the *maximum HBT overlap*, i.e. the length of the longest HBT overlap, between $\{a\,(n)\}_1^N$ and $\{b(n)\}_1^M$. If no HTB overlap exists, then we set $L_{\max} = 0$.

- **DEFINITION**

We define the *head-bites-tail (HBT) distance d* between $\{a\,(n)\}_1^N$ and $\{b(n)\}_1^M$ to be

$$d := d(a(n),\, b(n)) = N + M - 2\, L_{\max}$$

where $L_{\max}$ is the maximum HBT overlap between $a(n)$ and $b(n)$.

Intuition: d can also be thought of as specifying the number of remaining elements in $a(n)$ and $b(n)$ that DO NOT overlap.

- **Examples**

1. $\{a(n)\} = \qquad\qquad\qquad\qquad\qquad\qquad\qquad \{\mathbf{55}, 89, 144, 233, 377, 610\}$
$\quad\{b(n)\} = \{1, 1, 2, 3, 5, 8, 13, 21, 34, \mathbf{55}\}$
$d = 6 + 10 - 2\,(1) = 14$

2. $\{a(n)\} = \qquad\qquad \{\mathbf{3, 5, 8, 13, 21, 34, 55}, 89, 144, 233, 377\}$
$\quad\{b(n)\} = \{1, 1, 2, \mathbf{3, 5, 8, 13, 21, 34, 55}\}$
$d = 11 + 10 - 2\,(7) = 7$

Relative HBT Distance

- **DEFINITION**

We define the *relative HBT distance* d_r between $\{a\,(n)\}_1^N$ and $\{b(n)\}_1^M$ to be

$$d_r := d_r(a(n),\, b(n)) = \frac{d}{N+M} = \frac{N+M-2\,L}{N+M} = 1 - \frac{2\,L}{N+M}$$

NOTE: $0 \le d_r \le 1$

- **Examples**

1. $\{a(n)\} = \{\mathbf{55}, 89, 144, 233, 377, 610\}$
   $\{b(n)\} = \{1, 1, 2, 3, 5, 8, 13, 21, 34, \mathbf{55}\}$
$d_r = \frac{6+10-2\,(1)}{6+10} = \frac{14}{16} = \frac{7}{8}$

2. $\{a(n)\} = \{\mathbf{3, 5, 8, 13, 21, 34, 55}, 89, 144, 233, 377\}$
   $\{b(n)\} = \{1, 1, 2, \mathbf{3, 5, 8, 13, 21, 34, 55}\}$
$d_r = \frac{11+10-2\,(7)}{11+10} = \frac{7}{21} = \frac{1}{3}$

EUREKA Project

- **Implementation**

  i. *Mathematica* - generate sequence transformations and perform pattern matching $(d_r \leq 1/2, L_{\max} \geq 4)$

  ii. MySQL - store sequence transformations and matches to a database

- **Scope**

  i. First 170,000 sequences in OEIS (A000001-A170000)

  ii. Over one million sequence transformations (T1-T11)

- **Search Results**

  i. Over 300,000 matches found so far

  ii. Preliminary analysis shows:
  - Most matches are trivial or already mentioned in OEIS (> 99%)
  - Small fraction of false positives (> 0.9%)

Three Experimental Conjectures

- **EUREKA Database Website**

1. 1563: A000129S1T3 = A041011S1T8
2. 2010: A000240S1T7 = A006882S1T8
3. 2443: A000295S1T9 = A031878S1T4

Next Steps

- **Current Status**

  - Eureka Database contains more integer sequences than OEIS but not as “smart”

- **Scale up processing power and memory**

  - Perform search on a cluster of computers ✓

  - Implement parallel/distributed computing (Linux cluster)

- **Improve sequence matching algorithms**

  - Reduce search-times ✓

  - Reduce trivial matches and false positives

- **Expand Scope of Search**

  - Enlarge collection of sequence transformations ✓

  - Compositions of sequence transformations

  - Extend search to 2-D sequences (e.g. Pascal’s triangle) and rational sequences (e.g. Bernoulli numbers)

- **Disseminate Work**

  - Create database website ✓

  - Make database website accessible to the public

  - Publish new interesting (non-trivial) proofs of experimental conjectures

- **Seek Help**

  - Need good programmers (recruit students! ✓)

  - Need collaborators (faculty and students) to analyze and prove experimental conjectures (suitable as student research projects)

The End