УДК004.421.2:519.71

ИСПОЛЬЗОВАНИЕ МЕТОДА СЕКВЕНЦИАЛЬНОГО АНАЛИЗА ДЛЯ МОДЕЛИРОВАНИЯ ОБЪЕКТА С ФРАКТАЛЬНОЙ СТРУКТУРОЙ

КРАВЕЦ Н.С.

Показывается возможность использования алгоритма поиска ассоциативных правил, учитывающего фактор времени и взаимосвязь событий, для восстановления параметров модели детерминированной L-системы по последовательности символов.

Ключевые слова: фрактал, L-системы, секвенциальный анализ.

Key words: fractal, L-systems, sequential pattern mining.

1. Введение

Фрактал – нерегулярное самоподобное множество. Примеры самоподобных множеств известны с XIX века. Термин «фрактал» (от лат. fractus – раздробленный) впервые ввел в 1975 году математик исследовательского центра IBM Бенуа Мандельброт.

По способу построения выделяют фракталы детерминированные (геометрические и алгебраические) и недетерминированные (стохастические). Геометрические фракталы строятся на основе исходной фигуры путем ее дробления и выполнения различных преобразований полученных фрагментов. Алгебраические фракталы – строятся на основе алгебраических формул. Стохастическими называются фракталы, при построении которых случайным образом изменяются какие-либо параметры. Свойства фракталов:

– нетривиальная структура, которая не упрощается при увеличении масштаба;

– самоподобие;

– дробная метрическая размерность;

– возможность построения при помощи рекурсивной процедуры.

Многие природные объекты обладают фрактальными свойствами. Фракталы применяются в компьютерной науке (фрактальное сжатие данных, децентрализованные сети), физике (моделирование сложных процессов), биологии (моделирование популяций, сложных ветвящихся структур), технике (фрактальные антенны), экономике (фрактальный анализ временных рядов), социологии. В компьютерной графике фракталы используются для генерации изображений природных объектов. Актуальной является задача изучения фрактальных свойств различных объектов.

Целью работы является изучение возможности использования алгоритмов секвенциального анализа для поиска самоподобных элементов изображения или геометрической фигуры. Задачи, которые необходимо решить для этого: изучение моделей фракталов, построенных с помощью L-систем; выбор алгоритма поиска шаблонов последовательных событий (секвенциального анализа, sequential pattern mining или SPM), пригодного для анализа структуры геометрической фигуры с фрактальными свойствами.

2. Свойства L-систем

L-системы разработаны Аристидом Линденмайером, венгерским биологом-теоретиком и ботаником. Линденмайер использовал L-системы для описания поведения клеток растений, моделирования процессов роста и развития растений, морфологии различных организмов (рис.1). Рекурсивная природа правил L-системы ведет к самоподобию и тем самым фракталоподобные формы легко описываются L-системами [1].

Детерминированную контекстно-свободную L-систему определяют как кортеж:

$$G = (V, \omega, P),$$

где V – алфавит, набор символов; ω – аксиома или инициатор, строка символов из V, определяющая начальное состояние системы; P – набор правил, которые определяют, как переменные могут быть заменены комбинациями констант и других переменных.

Во время каждой итерации алгоритма каждый символ заменяется набором символов в соответствии с правилами. Если имеется ровно одно правило для каждого символа, то L-система называется детерминированной (детерминированная контекстно-свободная L-система, DOL-система). Если имеется несколько правил и каждое из них выбирается с определенной вероятностью при каждой итерации, L-система называется стохастической. Различные инструменты 3d моделирования используют L-системы для генерации контента.

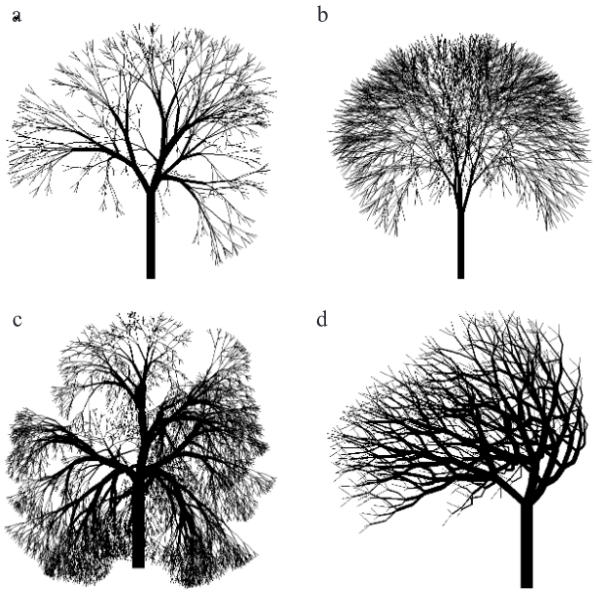

Рис. 1. Модели деревьев, построенные с использованием L-систем [2]

Рассмотрим решение противоположной задачи, а именно определение аксиом и правил L-системы по готовому изображению или его описанию.

Рассмотрим модель фрактала дракон Хартера-Хейтвея:

$$V = \{F, X, Y, +, -\}; \omega = FX; P = \{X \to X + YF; Y \to FX - Y\},$$

К символам алфавита добавим интерпретацию: F – двигаться вперед, нарисовать линию; X, Y – переменные; + – поворот на 90°; - – поворот на -90°.

В результате выполнения четырёх итераций алгоритма получаем следующие последовательности символов (табл. 1).

Таблица 1

№	Последовательность
0	FX
1	FX+YF
2	FX+YF+FX–YF
3	FX+YF+FX–YF+FX+YF–FX–YF
4	FX+YF+FX–YF+FX+YF–FX–YF+FX+YF+FX–YF–FX+YF–FX–YF

Изображение, соответствующее 4-му поколению последовательности, представлено на рис. 2.

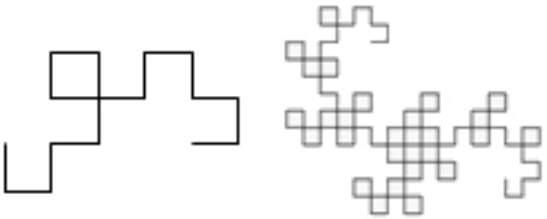

Рис. 2. Четвёртое и седьмое поколения дракона Хартера-Хейтвея

Изменение длины строки последовательности может быть представлено в виде ряда: $a_i = a_{i-1} \times 2 + 1, i = \overline{1,n}$.

Алгоритмы Data Mining по определению Пятницкого-Шапиро являются методами поиска в больших объемах данных неочевидных, объективных закономерностей. Рассмотрим возможность использования для построения модели подобного изображения в терминах L-систем алгоритмов формирования ассоциативных правил.

3. Алгоритмы SPM

Поиск ассоциативных правил предполагает нахождение частых зависимостей между объектами или событиями из большого числа наборов данных. Полученные ассоциации представляются в виде правил и могут использоваться как для лучшего понимания природы анализируемых данных, так и для прогнозирования. Однако при этом не учитывается такой атрибут транзакции как время.

Секвенциальный анализ (SPM) – разновидность задачи поиска ассоциативных правил. Целью SPM является установление отношения порядка между исследуемыми наборами. Наиболее распространенными алгоритмами SPM являются AprioriALL и GSP. Алгоритм AprioriALL был предложен в 1995 году (Argwal и Srikant). Он, также как другие усовершенствования Apriori, основывается на утверждении, что последовательность, входящая в часто встречающуюся последовательность, также является часто встречающейся.

Исходными данными для алгоритма AprioriALL является упорядоченный набор событий.

Пусть изображение L-системы (модель геометрического фрактала) представлено в виде ломаной линии, заданной последовательностью вершин вида: $\langle P_1(x_1, y_1), .., P_n(x_n, y_n) \rangle$, где $n \geq 3, n \in Z$. Чтобы получить последовательность символов, аналогичную L-системе, необходимо заменить каждый отрезок P_iP_{i+1} из набора $P_1P_2, P_2P_3, .., P_{n-2}P_{n-1}$ последовательностью $\{F\}...\{X\}...$, где $X = \begin{cases} "-", (y_i - y_{i+1}) < (y_{i+2} - y_{i+1}), \\ "+", (y_i - y_{i+1}) > (y_{i+2} - y_{i+1}). \end{cases}$ F — единичный отрезок, длина которого равна длине самого короткого отрезка исходной ломаной; «+» или «–» соответствуют углу поворота ломаной против часовой стрелки или по часовой стрелке; размер угла вычисляется как угол α_i между векторами $\overline{P_iP_{i+1}}$ и $\overline{P_{i+1}P_{i+2}}$; количество повторений «+» или «–» равно $\frac{\alpha_i}{\alpha_{min}}$, где $\alpha_1, \alpha_2, .., \alpha_{n-2}$ — углы между отрезками ломаной $P_1P_2..P_n$. Отрезок $P_{n-1}P_n$ заменяется только символом F.

Представим в виде таблицы (табл. 2) данные о последовательности, описывающей L-систему в виде, соответствующем требованиям алгоритма AprioriALL.

Таблица 2

AprioriALL	L-система
D — множество всех транзакций T;	строка символов, например, строка F-F++F-F++F-F++F-F++F-F++F-F;
T — транзакция, заданная как (ID, дата/время, список объектов);	ID, дата/время определяются номером транзакции, список объектов — это символ строки или несколько одинаковых подряд идущих символов;
I — множество всех объектов (товаров);	множество символов алфавита, определяемое всеми различными символами, входящими в строку, например, {F, +, –};
s_i — набор, состоящий из элементов I;	набор символов, входящих в транзакцию;
S — последовательность, упорядоченный список наборов.	последовательность символов, входящих в строку, например, F+.

Последовательность символов будем считать клиентской, т.е. соответствующей набору всех транзакций одного клиента, упорядоченных по времени. Периоды времени между любыми двумя транзакциями T_i и T_{i+1} будем считать равными.

Поддержкой последовательности S называется отношение количества транзакций, в которое входит последовательность S, к общему количеству транзакций. Задачей SPM является поиск всех частых последовательностей, т.е. тех, для которых уровень поддержки находится в пределах, обозначенных минимальным и максимальным значением. Оба параметра задаются пользователем до начала работы алгоритма.

Работа алгоритма состоит из нескольких фаз: сортировки; отбора кандидатов; трансформации; генерации последовательностей; максимизации. Ограничением алгоритма AprioriAll является отсутствие возможности определения силы взаимосвязи, т.е. промежутка времени, разделяющего события.

В алгоритме GSP (Generalized Sequential Pattern algorithm) введены дополнительные параметры, благодаря чему стало возможным учитывать ограничения по времени между соседними транзакциями последовательности. В общем виде работа алгоритма GSP похожа на AprioriAll, при этом, как указано в [3], GSP работает быстрее почти в 20 раз. Возможность учитывать ограничения по времени важна для поиска самоподобных элементов в L-системах. Повторное применение такого алгоритма к строке L-системы и сравнение результатов позволит восстановить набор правил P, задающих L-систему.

4. Алгоритм GSP

Рассмотрим подробнее работу алгоритма GSP. Каждая итерация алгоритма предусматривает «проход» по исходному набору данных. Во время первой итерации вычисляется поддержка для каждого объекта (одноэлементной последовательности) и выполняется фильтрация. В результате исходные данные для следующей итерации алгоритма формируются из последовательностей, чья поддержка равна либо превышает пороговое значение. Далее генерируются более длинные последовательности-кандидаты, снова подсчитывается их поддержка и снова производится фильтрация, результаты которой послужат исходными данными для следующего шага алгоритма. Число элементов последовательностей–кандидатов на каждой итерации алгоритма одинаково.

К дополнительным параметрам алгоритма относятся: минимальное и максимальное допустимое время между транзакциями (min-gap и max-gap); размер скользящего окна (w), интервала времени, в пределах которого наборы объектов из одного элемента последовательности могут принадлежать разным транзакциям. Для заданных параметров последовательность $d = \langle d_1..d_m \rangle$ содержит последовательность $s = \langle s_1..s_m \rangle$, если существуют такие целые числа $l_1 \le u_1 \le l_2 \le u_2 < ... < l_n \le u_n$, что:

а) s_i содержится в $\bigcup_{k=l_i}^{u_i} d_k, 1 \le i \le n$;

б) для периода времени между транзакциями d_{u_i} и d_{l_i}, Δ выполняются следующие условия:

$$\Delta \le w, 1 \le i \le n;$$

$$\Delta > min-gap, 2 \le i \le n;$$

$$\Delta \le max-gap, 2 \le i \le n.$$

Генерация последовательностей-кандидатов производится объединением последовательностей меньшего размера. Объединение последовательностей s_i и $s_j, j > i$ происходит путём добавления к s_i последнего элемента из s_j, если s_j содержит смежную подпоследовательность s_i. Последовательность c является смежной подпоследовательностью $s = \langle s_1 s_2 .. s_n \rangle$, если:

а) c получается из s при удалении элемента s_1 или s_n;

б) c получается из s при удалении одного объекта из элемента s_i, если в его составе не менее двух объектов;

в) c – смежная подпоследовательность c', если c' – смежная подпоследовательность s.

Работа алгоритма завершается тогда, когда не найдено ни одной новой последовательности с достаточным уровнем поддержки в конце очередного шага или когда невозможно сформировать новых кандидатов.

Выводы

Обоснована возможность использования алгоритмов SPM для анализа структуры геометрической фигуры с фрактальными свойствами, представленной в терминах L-систем. Для данной задачи лучше подходит алгоритм GSP благодаря его возможности учитывать минимальное и максимальное время между транзакциями. Дальнейшей проработки требуют проблемы восстановления правил L-системы, содержащих переменные, символы, не имеющие геометрической интерпретации, а также L-систем, состоящих из нескольких ломаных линий.

Литература: 1. *Capasso V.* et al. (ed.). Pattern Formation in Morphogenesis: problems and mathematical issues. – Springer Science & Business Media, 2012. T. 15. P. 137-151. **2.** *Prusinkiewicz P., Lindenmayer A.* The algorithmic beauty of plants. Springer Science & Business Media, 2012. **3**. *Srikant R., Agrawal R.* Mining sequential patterns: Generalizations and performance improvements. Springer Berlin Heidelberg, 1996. P. 1-17.

Транслитерированный список литературы. 1. *Capasso V.* et al. (ed.). Pattern Formation in Morphogenesis: problems and mathematical issues. Springer Science & Business Media, 2012. T. 15. P. 137-151. **2.** *Prusinkiewicz P., Lindenmayer A.* The algorithmic beauty of plants. Springer Science & Business Media, 2012. 3. *Srikant R., Agrawal R.* Mining sequential patterns: Generalizations and performance improvements. Springer Berlin Heidelberg, 1996. P. 1-17.

Поступила в редколлегию 02.12.2015

Рецензент: д-р техн. наук, проф. Кириченко Л.О.

Кравец Наталья Сергеевна, канд. техн. наук, доцент кафедры информационно-документных систем Харьковской государственной академии культуры. Научные интересы: моделирование информационных систем, Data Mining. Адрес: Украина, 61057, Харьков, Бурсацкий спуск, 4, тел. (057) 731-32-82. E-mail:kravets_n@list.ru.