점집합의 최대 반복 패턴

Véronique Cortier Xavier Goaoc 이미라° 나현숙1)
INRIA INRIA 한국과학기술원 숭실대학교
{cortier, goaoc}@loria.fr mira@kaist.ac.kr hsnaassu.ac.kr

Maximally repeated sub-patterns of a point set

Véronique Cortier Xavier Goaoc Mira Lee° Hyeon-Suk Na
INRIA(France) INRIA(France) KAIST Soongsil Univ.

요 약

We answer a question raised by P. Brass on the number of maximally repeated sub-patterns in a set of n points in $\mathbb{R}^d$. We show that this number, which was conjectured to be polynomial, is in fact $\Theta(2^{\frac{n}{2}})$ in the worst case, regardless of the dimension d.

1. Introduction

Let S be a set of n points in $\mathbb{R}^d$. A *sub-pattern*, i.e. a subset, of S is repeated if it can be translated to another subset of S. A sub-pattern $P \subseteq S$ is *maximally repeated* if for any subset Q such that $P \subseteq Q \subseteq S$ there exists a translation that maps P to a subset of S without mapping Q to a subset of S. In other words, a pattern is maximally repeated if it cannot be extended without losing at least one of its occurrences.

Maximally repeated sub-patterns (MRSP for short) originated from the field of pattern matching to solve the following problem: given two point sets X and Y, can Y be translated to a subset of X? P. Brass[Theorem 3 in [1]] gave an algorithm that answer such queries in time $O(|Y|\log|X|)$ whose preprocessing time depends on the number of distinct MRSP of X, where two MRSP are *distinct* if they are not equal up to a translation. A natural question is thus to give a theoretical bound on this number of MRSP in order to provide an upper bound on the time requirement of that algorithm. This number was conjectured [[1] or page 267 in [2]] to be $O(n^d)$ where d is the dimension in which the point set is embedded.

In this note we show that the number of MRSP of a set of n points is actually $\Theta(2^{\frac{n}{2}})$ in the worst case, which shows that finding sub-patterns via this approach may lead to exponential running time in the worst-case. Our proof is based on combinatorial rather than geometrical properties of the point set, which explains that the bound is independent of the dimension d in which the points are considered.

2. Lower and Upper bounds

Let us first introduce some terminology. Given $P \subseteq \mathbb{R}^d$ and $t \in \mathbb{R}^d$, the *translation* of P by t, denoted $P+t$, is the set $\{x+t \mid x \in P\}$. A subset $P \subseteq S$ is a repeated sub-pattern if there exists a translation $t \neq 0$ such that $P+t \subseteq S$. P is a *maximally repeated sub-pattern* (MRSP) if, in addition, for any subset Q such that $P \subseteq Q \subseteq S$ there exists a translation t such that $P+t \subseteq S$ and $Q+t \not\subseteq S$. Two MRSP are *distinct* if they are not

1) This work was supported by the Korea Research Foundation Grant.. (KRF-2004-000-10004-0)

equal up to a translation.

2.1. Lower bound

We build our example on a 1-dimensional grid which can, of course, be considered as embedded in $\mathbb{R}^d$ for any $d \geq 1$. Let k be an integer, G_k denotes the set of integers $\{1, \ldots, k\}$ and $S_k = G_k \cup (G_k + k + 1)$, that is two copies of G_k separated by a gap of one point at $k+1$.

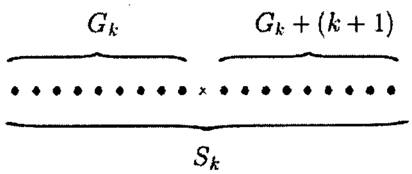

그림 1: Example for Lower Bound 2^{k-1}

Lemma 1
The set S_k has 2^{k-1} distinct MRSP.

Proof.
We show that any subset $P \subseteq G_k$ is a MRSP by arguing that for any point $p^* \in S_k/P$, one of the translations that keeps P in S_k sends p^* either to $\{k+1\}$ or outside of S_k. Indeed, let $Q \subseteq S_k$ be a proper superset of P and $t \in Q/P$. If $t \geq k+2$ then $P + (k+1) \subseteq S_k$ and $Q + (k+1) \not\subseteq S_k$.
If $t \leq k$ then $P + (k+1-t) \subseteq S_k$ and $Q + (k+1-t) \not\subseteq S_k$. This proves that any subset $P \subseteq G_k$ is a MRSP of S_k. No translation can map a subset of G_k that contain 1 to another subset of G_k that contains 1. Therefore, at least 2^{k-1} of the subsets of G_k are distinct MRSP. ◇

2..2 Upper bound

Let $S = \{a_1, \ldots, a_n\} \subseteq \mathbb{R}^d$ be a set of n points. We consider the *set of translations* T defined by

$$T = S - S = \{x - y \mid (x, y) \in S^2\}$$

Both the points in S and the vectors in T are ordered lexicographically, as vectors of n real numbers. Let A denote the family of all *first* occurrences of subsets of S that are MRSP. By "first" we mean that a MRSP P is in A if and only if no translation $t < 0$ satisfies $P + t \subseteq S$. That is, we choose one representative of each equivalence class of MRSP under translation. The following function maps each pattern to its set of translations:

$$\phi : \begin{array}{l} 2^S \rightarrow 2^T \\ P \mapsto \{t \in T : P + t \subseteq S\} \end{array}$$

For $1 \leq i \leq j \leq n$, let

$$A_{ij} = \{P \in A : \{a_i, a_j\} \subseteq P \subseteq \{a_i, \ldots, a_j\}$$

be the set of all occurrences of MRSP spanning the range $\{a_i, \ldots, a_j\}$ and

$$T_{ij} = \{t \in T : t \geq 0, \{a_i, a_j\} \subseteq S \cap (S - t)\}$$

be the set of all non-negative translations compatible with a_i and a_j. We can now prove our upper bound.

Lemma 2
A set of n points has at most $16 \cdot 2^{n/2}$ distinct MRSP.

Proof.
Let P_1 and P_2 be two MRSP such that $\phi(P_1) = \phi(P_2)$. Then $\phi(P_1 \cup P_2) = \phi(P_1) = \phi(P_2)$ which leads to $P_1 \cup P_2 = P_1$, since P_1 is a MRSP, and $P_1 \cup P_2 = P_1$, as P_2 is also a MRSP. Thus, ϕ defines an injection from A on the subsets of T. If $P \in A_{ij}$ then $\phi(P) \subseteq T_{ij}$ and ϕ induces an injection from A_{ij} on the subsets of T_{ij}. Hence, $A_{ij} \leq 2^{|T_{ij}|}$.

If $t \in T_{ij}/\{0\}$ then $t > 0$ and $a_j + t = a_y$ with $y > j$. Hence, $|T_{ij} - \{0\}| \leq n - j$. It follows that

$$|A_{ij}| \leq 2^{n-j} - 1 .$$

As any MRSP in A_{ij} corresponds to a subset of

$\{i+1, \ldots, j-1\}$, we also have that

$$|A_{ij}| \leq 2^{j-i-1}.$$

Note that A_{ij} is empty for $i \geq 2$ and A_{11} is a singleton. We can now write

$$|A| \leq 1 + \sum_{i=1}^{n} \sum_{j=i+1}^{n} 2^{min(n-j, j-i-1)}$$

Splitting the sum at $j = \frac{n+i}{2} + 1$, we get

$$|A| \leq 1 + 2\sum_{i=1}^{n} \sum_{j=i+1}^{\frac{n+i}{2}} 2^{j-i-1} = 1 + 2\sum_{i=1}^{n} 2^{\frac{n-i}{2}+1}$$

$$\leq 1 + 8\sum_{l=0}^{\frac{n}{2}} 2^{l} \leq 16 \cdot 2^{n/2} \diamond$$

[Reference]

[1] P. Brass. Combinatorial geometry problems in pattern recognition. *Discrete and Computational Geometry*, 28: 495-510, 2002.
[2] P. Brass, W. Moser, and J. Pach. *Research Problems in Discrete Geometry*, Springer-Verlag, 2005.

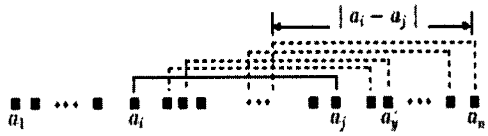

그림 2. Bounding $|T_{ij}|$ in 1-dimensional case; the same reasoning holds in $\mathbb{R}^d$ thanks to the total ordering.