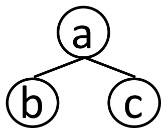
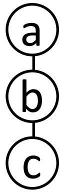
Algo+index

6.12

transformation

- Normally, the tree can be encoded into one sequence, however, one sequence does not necessarily correspond to one unique tree.
- E.g., {a, b, c} can be converted by





- In our scenario:
  - Each node in P-tree is unique, which means the position of each node is fixed.
  - Thus one P-tree one-to-one corresponds to one sequence.
  - A sequence $< a_1, a_2 \cdots a_n >$ is a subsequence of another sequence $< b_1, b_2 \cdots b_m >$ if there exists integers $i_1 < i_2 < \cdots < i_n$ such that $a_1 = b_{i_1}, a_2 = b_{i_2} \cdots a_n = b_{i_n}$. ($m > n$)
  - *Maximal common subtree* $\longleftrightarrow$ *maximal common subsequence*

[1]Agrawal R, Srikant R. Mining sequential patterns[C]//Data Engineering, 1995. Proceedings of the Eleventh International Conference on. IEEE, 1995: 3-14.

Maximal common subsequence(MCS)

- SIGKDD-04[1]: ”Maximal frequent subsequences ”(MFS) is NP-hard.
- MCS has two similar steps to MFS: enumeration and counting.

- In our scenario:
  - One query node(corresponds to one unique sequence S of its P-tree) is required which is not defined in MFS.
  - However, in worst case, S is encoded by the complete taxonomy.
  - Thus, MCS is NP-hard.

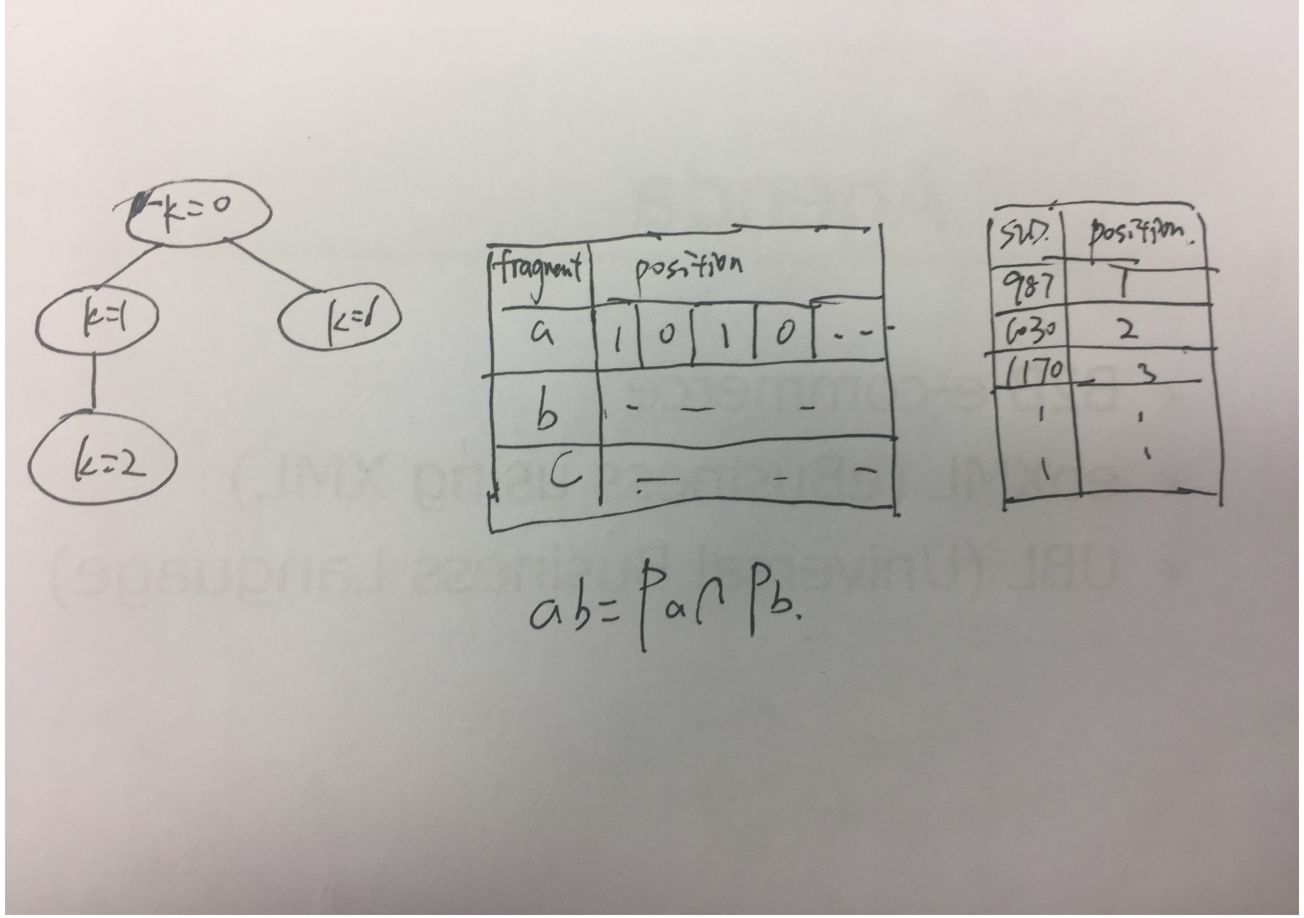
[1] Yang G. The complexity of mining maximal frequent itemsets and maximal frequent patterns[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004: 344-353.

basic algorithm Ⅱ

- Longest common subsequence(lcs) between two sequence is solvable in polynomial time by dynamic programming.
- Lemma: sequences with same length are not subsequence of each other.

- Basic algorithm Steps:
  - K-$\widehat{core}$ search for G'.
  - For each $v \in G'$, $l_i = l_i \cup$ lcs(v, q). i represents the length.
  - Mine MCS. (anti-monotonicity or hash-tree)
  - Recheck connectivity and etc.

Naïve Index



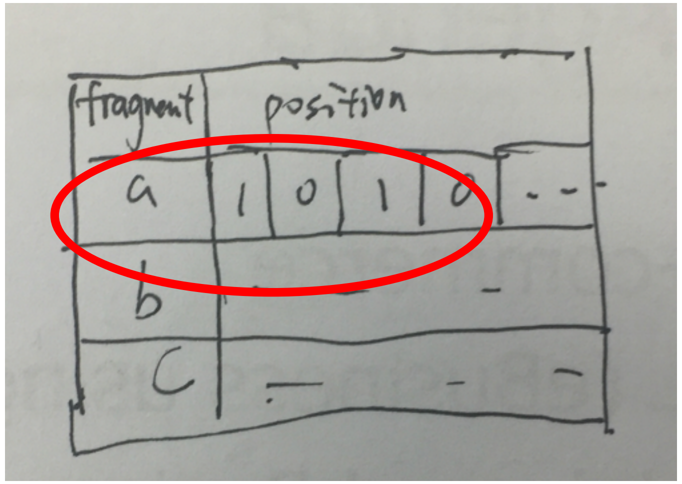
k=0
k=1
k=1
k=2
fragment
position
a
1
0
1
0
b
c
SID.
position.
987
1
630
2
1170
3
ab=Pa∩Pb.

compressed index-mathematic preliminary

- The only divisors of a prime integer p (p>1) are 1 and p.
- Every positive integer n is either 1 or can be expressed as a product of several prime integers, and this factorization is unique with the order of prime integers. The *standard form* of n factorization of n: n=${p_1}^{m_1}{p_2}^{m_2}\cdots {p_n}^{m_n}$, p_i is a distinct prime integer, m_i is called the *multiplicity* of p_i.
- Give two integers a and b, the great common divisor of a, b is $\gcd(a, b)$. E.g., a =$2^3 \cdot 3^2 \cdot 7 = 504$, $b = 2^2 \cdot 3^1 \cdot 7 \cdot 11$ =924, $\gcd(a, b) = 2^2 \cdot 3^1 \cdot 7 = 84$.
- if we set m_i as 1. Then n=$p_1 p_2 \cdots p_n$. $\gcd(a, b)$=$\prod_{i=1}^{m} p_{x_i}$.

compressed index-mathematic preliminary

- $G = \{2,3,5,7\}$, $a = 2 \cdot 3 = 6$, $b = 5 \cdot 7$=35, gcd(a,b)=1.
- $S_a = 1100$, $S_b = 0011$, $S_a \cap S_b$=0000=1.



- Set 4 bits as a block.S_a={10 ,⋯}.

Then $S_a \cap S_b$={gcd(S_{a_i}, S_{b_i})| $i \in number\ of\ blocks$}.

- If G={2,3,5,7}, then gcd(a,b) has $2^4 \cdot 2^4 \cdot 0.5$=128 types which can be pre-computed and stored in a table.

Compressed index