

Hypersistent Sketch: Enhanced Persistence Estimation in Web Data Stream via Rapid Item Separation

Algorithm 1: Insert_stage1

Input: Item e
Output: Report whether the insertion of element e in stage1 was successful

```

1 for  $entry \in L0[hash0(e)]$  %  $bucket\_num$  do
2   if  $entry.id == null \parallel entry.id == id$  then
3      $entry.id = e.id$ ;
4     return true;
5   end
6 endfor
7 return false;
  
```

Algorithm 2: Insert_stage2

Input: Item e
Output: Report whether the insertion of element e in stage2 was successful

```

1  $v1 = \min_{1 \leq i \leq d1} L1[i][hash1_i(e)].cnt$ ;
2 if  $v1 < threshold_{L1}$  then
3   for  $i$  from 1  $\rightarrow d1$  do
4      $b_1 \leftarrow L1[i][hash1_i(e)]$ ;
5     if  $b_1.cnt == v1 \ \&\& \ b_1.flag == On$  then
6        $b_1.cnt \leftarrow b_1.cnt + 1$ ;
7        $b_1.flag \leftarrow Off$ ;
8     end
9   endfor
10  return true;
11 end
12                                      $\triangleright L1$  reach threshold
13  $v2 = \min_{1 \leq i \leq d2} L2[i][hash2_i(e)].cnt$ ;
14 if  $v2 < threshold_{L2}$  then
15   for  $i$  from 1  $\rightarrow d2$  do
16      $b_2 \leftarrow L2[i][hash2_i(e)]$ ;
17     if  $b_2.cnt == v2 \ \&\& \ b_2.flag == On$  then
18        $b_2.cnt \leftarrow b_2.cnt + 1$ ;
19        $b_2.flag \leftarrow Off$ ;
20     end
21   endfor
22  return true;
23 end
24                                      $\triangleright L2$  reach threshold
25 return false;
  
```

Algorithm 3: Insert_stage3

Input: Item e

```
1  $h_3 \leftarrow \text{hash3}(e) \% (\text{length} * (\text{length} - 1));$ 
2  $h_{3\_1} \leftarrow h_3 \% \text{length}; h_{3\_2} \leftarrow h_3 \div (\text{length} - 1);$ 
3  $\text{replace} = \text{null};$ 
4 for  $i \in \{1, 2\}$  do
5    $b_3 \leftarrow L3[i][h_{3\_i}];$ 
6   for  $\text{entry} \in b_3$  do
7     if
8        $\text{entry.item} == \text{null} \parallel \text{entry.item} == e \ \&\& \ \text{entry.flag} == \text{On}$ 
9       then
10         $\text{entry.item} \leftarrow e;$ 
11         $\text{entry.flag} \leftarrow \text{Off};$ 
12         $\text{entry.cnt} \leftarrow \text{entry.cnt} + 1;$ 
13        return;
14      end
15    else
16      if  $\text{replace} == \text{null} \parallel \text{entry.cnt} < \text{replace.cnt}$  then
17         $\text{replace} \leftarrow \text{entry};$ 
18      end
19    end
20  endfor
21 endfor
22 if  $h_3 \% (\text{replace.cnt} + 1) == 0$  then
23    $\text{replace.item} \leftarrow e;$ 
24   if  $\text{replace.flag} == \text{On}$  then
25      $\text{replace.cnt} \leftarrow \text{replace.cnt} + 1;$ 
26      $\text{replace.flag} \leftarrow \text{Off};$ 
27   end
28 end
```

Algorithm 4: flush_L0

```
1   ▷ when new window arrives, flush items in  $L0$  to the later stage
2 for  $\text{entry} \in L0$  do
3   if  $\text{entry} \neq \text{null}$  then
4     if  $\text{!Insert\_stage2}(e)$  then
5        $\text{Insert\_stage3}(e);$ 
6     end
7   end
8 endfor
```

Algorithm 5: Insert

Input: Item e

```
1 if  $Insert\_stage1(e)$  then
2   | return;
3 end
4 if  $Insert\_stage2(e)$  then
5   | return;
6 end
7  $Insert\_stage3(e)$ ;
8 return;
```

Algorithm 6: Query

Input: Item e

Output: The persistence of e

```
1  $ret \leftarrow 0$ ;
2  $v1 \leftarrow \min_{1 \leq i \leq d1} (L1[i][hash1_i(e)])$ ;
3 if  $v1 < threshold_{L1}$  then
4   | return  $v1$ ;
5 end
6  $ret \leftarrow ret + v1$ ;
7  $v2 = \min_{1 \leq i \leq d2} (L2[i][hash2_i(e)])$ ;
8 if  $v2 < threshold_{L2}$  then
9   | return  $ret + v2$ ;
10 end
11  $ret \leftarrow ret + v2$ ;
12  $hash\_num \leftarrow hash3(e) \% length * (length - 1)$ ;
13  $bucket\_pos_1 \leftarrow hash\_num \% length$ ;
14  $bucket\_pos_2 \leftarrow hash\_num \div (length - 1)$ ;
15 for  $i \in \{1, 2\}$  do
16   |  $bucket \leftarrow L3[i][bucket\_pos_i]$ ;
17   | for  $entry \in bucket$  do
18     | if  $entry.item == e$  then
19       |  $ret \leftarrow ret + entry.count$ ;
20     | end
21   | endfor
22 endfor
23 return  $ret$ ;
```

Algorithm 7: Scanning a bucket with 16 cells through the SIMD instructions

Input: The incoming item e and the start address p of the bucket

Output: Return the index of the matched key or -1

```
1 __m128i item = _mm_set1_epi32(e);
2 __m128i * keys_p = (__m128i *) p;
3 __m128i a_comp = _mm_cmpeq_epi32(item, keys_p[0]);
4 __m128i b_comp = _mm_cmpeq_epi32(item, keys_p[1]);
5 __m128i c_comp = _mm_cmpeq_epi32(item, keys_p[2]);
6 __m128i d_comp = _mm_cmpeq_epi32(item, keys_p[3]);
7 a_comp = _mm_packs_epi32(a_comp, b_comp);
8 c_comp = _mm_packs_epi32(c_comp, d_comp);
9 a_comp = _mm_packs_epi32(a_comp, c_comp);
10 matched = _mm_movemask_epi8(a_comp);
11 if match  $\neq$  0 then
12   | return TZCNT(matched);
13 end
14 else
15   | return -1;
16 end
```
