

The background features a complex network of thin, light-colored lines forming a web-like structure. Scattered throughout are numerous small, colored dots in shades of green, blue, and orange. A prominent, thicker red line forms a large, irregular loop in the center. The overall aesthetic is technical and data-driven.

GSP: Apriori-Based Sequential Pattern Mining



GSP: Apriori-Based Sequential Pattern Mining

- Initial candidates: All singleton sequences
 - <a>, , <c>, <d>, <e>, <f>, <g>, <h>
- Scan DB once, count support for each candidate
- Generate length-2 candidate sequences

$min_sup = 2$

| Cand. | sup |
|-------|-----|
| <a> | 3 |
| | 5 |
| <c> | 4 |
| <d> | 3 |
| <e> | 3 |
| <f> | 2 |
| <g> | 1 |
| <h> | 1 |

| | <a> | | <c> | <d> | <e> | <f> |
|-----|------|------|------|------|------|------|
| <a> | <aa> | <ab> | <ac> | <ad> | <ae> | <af> |
| | <ba> | <bb> | <bc> | <bd> | <be> | <bf> |
| <c> | <ca> | <cb> | <cc> | <cd> | <ce> | <cf> |
| <d> | <da> | <db> | <dc> | <dd> | <de> | <df> |
| <e> | <ea> | <eb> | <ec> | <ed> | <ee> | <ef> |
| <f> | <fa> | <fb> | <fc> | <fd> | <fe> | <ff> |

| | <a> | | <c> | <d> | <e> | <f> |
|-----|-----|--------|--------|--------|--------|--------|
| <a> | | <(ab)> | <(ac)> | <(ad)> | <(ae)> | <(af)> |
| | | | <(bc)> | <(bd)> | <(be)> | <(bf)> |
| <c> | | | | <(cd)> | <(ce)> | <(cf)> |
| <d> | | | | | <(de)> | <(df)> |
| <e> | | | | | | <(ef)> |
| <f> | | | | | | |

| SID | Sequence |
|-----|-----------------|
| 10 | <(bd)cb(ac)> |
| 20 | <(bf)(ce)b(fg)> |
| 30 | <(ah)(bf)abf> |
| 40 | <(be)(ce)d> |
| 50 | <a(bd)bcb(ade)> |

- Length-2 candidates:
 $36 + 15 = 51$
- Without Apriori pruning:
 $8 * 8 + 8 * 7 / 2 = 92$ candidates

GSP (Generalized Sequential Patterns): Srikant & Agrawal @ EDBT'96)

GSP Mining and Pruning

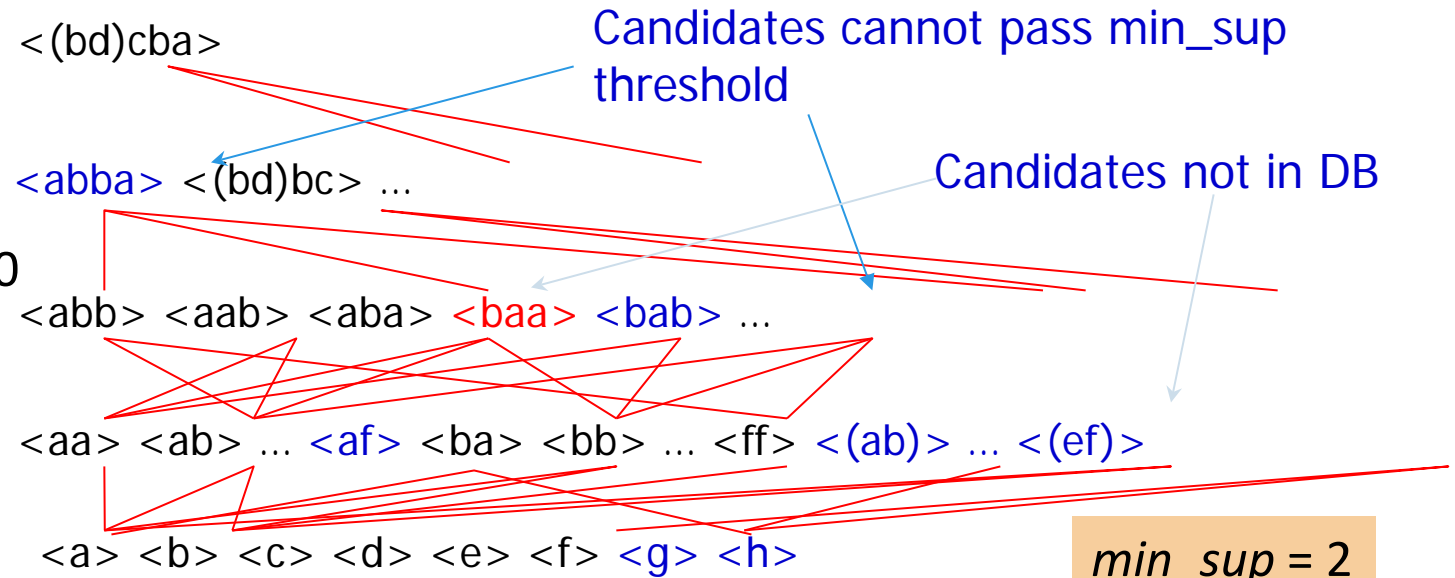
5th scan: 1 cand. 1 length-5 seq. pat.

4th scan: 8 cand. 7 length-4 seq. pat.

3rd scan: 46 cand. 20 length-3 seq. pat. 20 cand. not in DB at all

2nd scan: 51 cand. 19 length-2 seq. pat. 10 cand. not in DB at all

1st scan: 8 cand. 6 length-1 seq. pat.



- Repeat (for each level (i.e., length-k))
 - Scan DB to find length-k frequent sequences
 - Generate length-(k+1) candidate sequences from length-k frequent sequences using Apriori
 - set $k = k+1$
- Until no frequent sequence or no candidate can be found

| min_sup = 2 | |
|-------------|-----------------|
| SID | Sequence |
| 10 | <(bd)cb(ac)> |
| 20 | <(bf)(ce)b(fg)> |
| 30 | <(ah)(bf)abf> |
| 40 | <(be)(ce)d> |
| 50 | <a(bd)bcb(ade)> |