SS-ZG548: ADVANCED DATA MINING

Lecture-14: Clustering on Data Stream, Big Data

Dr. Kamlesh Tiwari
Assistant Professor
Department of Computer Science and Information Systems Engineering,
BITS Pilani, Rajasthan-333031 INDIA

March 24, 2018 (WILP @ BITS-Pilani Jan-Apr 2018)

Sequence Data

- Sequence Data: $S =< e_1, e_2, e_3, ... >$ attributed with specific time
- Each element e_i is a list of events $\{i_1, i_2, ..., i_k\}$
- Subsequence: $< a_1, a_2, ..., a_n >$ is contained in $< b_1, b_2, ...b_m >$ if $\exists i_1 < i_2 < i_3 < ... < i_n$ such that $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}$, ... , $a_n \subseteq b_{i_n}$
- Support for a sequence in database is the fraction that contains it

Sequence Data

- Sequence Data: $S =< e_1, e_2, e_3, ... >$ attributed with specific time
- Each element e_i is a list of events $\{i_1, i_2, ..., i_k\}$
- Subsequence: $< a_1, a_2, ..., a_n >$ is contained in $< b_1, b_2, ...b_m >$ if $\exists i_1 < i_2 < i_3 < ... < i_n$ such that $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}$, ... , $a_n \subseteq b_{i_n}$
- Support for a sequence in database is the fraction that contains it

Frequent sequence have support $\geq$*minsup*

Sequence Data

- Sequence Data: $S =< e_1, e_2, e_3, ... >$ attributed with specific time
- Each element e_i is a list of events $\{i_1, i_2, ..., i_k\}$
- Subsequence: $< a_1, a_2, ..., a_n >$ is contained in $< b_1, b_2, ...b_m >$ if $\exists i_1 < i_2 < i_3 < ... < i_n$ such that $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}$, ... , $a_n \subseteq b_{i_n}$
- Support for a sequence in database is the fraction that contains it

Frequent sequence have support $\geq$*minsup*

Consider sequences
$A =< \{1, 2, 4\}, \{2, 3\}, \{5\} >$
$B =< \{1, 2\}, \{2, 3, 4\} >$
$C =< \{1, 2\}, \{2, 3, 4\}, \{2, 4, 5\} >$
$D =< \{2\}, \{3, 4\}, \{4, 5\} >$
$E =< \{1, 3\}, \{2, 4, 5\} >$

Sequence Data

- Sequence Data: $S = < e_1, e_2, e_3, ... >$ attributed with specific time
- Each element e_i is a list of events $\{i_1, i_2, ..., i_k\}$
- Subsequence: $< a_1, a_2, ..., a_n >$ is contained in $< b_1, b_2, ...b_m >$ if $\exists i_1 < i_2 < i_3 < ... < i_n$ such that $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}$, ... , $a_n \subseteq b_{i_n}$
- Support for a sequence in database is the fraction that contains it

Frequent sequence have support $\geq$*minsup*

Consider sequences

$A = < \{1, 2, 4\}, \{2, 3\}, \{5\} >$
$B = < \{1, 2\}, \{2, 3, 4\} >$
$C = < \{1, 2\}, \{2, 3, 4\}, \{2, 4, 5\} >$
$D = < \{2\}, \{3, 4\}, \{4, 5\} >$
$E = < \{1, 3\}, \{2, 4, 5\} >$

Sequence	Support
$< \{1, 2\} >$	

Sequence Data

- Sequence Data: $S =< e_1, e_2, e_3, ... >$ attributed with specific time
- Each element e_i is a list of events $\{i_1, i_2, ..., i_k\}$
- Subsequence: $< a_1, a_2, ..., a_n >$ is contained in $< b_1, b_2, ...b_m >$ if $\exists i_1 < i_2 < i_3 < ... < i_n$ such that $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}$, **...** , $a_n \subseteq b_{i_n}$
- Support for a sequence in database is the fraction that contains it

Frequent sequence have support $\geq$*minsup*

Consider sequences

$A =< \{1, 2, 4\}, \{2, 3\}, \{5\} >$
$B =< \{1, 2\}, \{2, 3, 4\} >$
$C =< \{1, 2\}, \{2, 3, 4\}, \{2, 4, 5\} >$
$D =< \{2\}, \{3, 4\}, \{4, 5\} >$
$E =< \{1, 3\}, \{2, 4, 5\} >$

Sequence	**Support**
$< \{1, 2\} >$	60%
$< \{2, 4\} >$	80%
$< \{1\}, \{2\} >$	80%
$< \{1, 2\}, \{2, 3\} >$	60%

Generalized Sequential Pattern (GSP)[1]

S-1 First pass to yield all 1-element frequent sequences

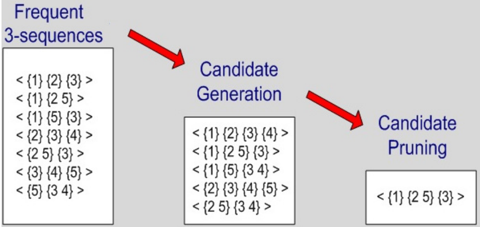
S-2 Repeat until new frequent sequences are found

- **Candidate Generation:** merge pairs found in $k-1^{th}$ pass. w_1 and w_2 can be merged if subsequences obtained by removal of first element of w_1 and last element of w_2 are same
- **Candidate Pruning:** Prune candidates that contain a subsequence which is infrequent in $k-1$ subseqeuces
- **Support Counting:** Need new pass to database
- **Candidate Elimination:** Involves thresholding based on minsup

$<\{1\},\{2,3\}\{4\}>$ and $<\{2,3\},\{4,5\}>$	$<\{1\},\{2,3\},\{4,5\}>$
$<\{1\},\{2,3\}\{4\}>$ and $<\{2,3\},\{4\},\{5\}>$	$<\{1\},\{2,3\},\{4\},\{5\}>$
$<\{1\},\{2,6\}\{4\}>$ and $<\{1\},\{2,6\},\{4\}>$	Can not be merged

[1] Generalized Sequential Pattern (GSP), Srikant and Agrawal, In EDBT 1996

Recap: Pruning in GSP



Frequent 3-sequences

< {1} {2} {3} >
< {1} {2 5} >
< {1} {5} {3} >
< {2} {3} {4} >
< {2 5} {3} >
< {3} {4} {5} >
< {5} {3 4} >

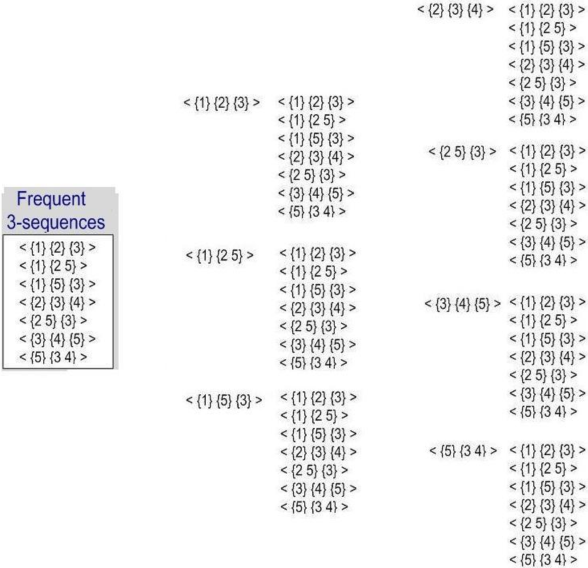
Candidate Generation

< {1} {2} {3} {4} >
< {1} {2 5} {3} >
< {1} {5} {3 4} >
< {2} {3} {4} {5} >
< {2 5} {3 4} >

Candidate Pruning

< {1} {2 5} {3} >

Candidate Generation



Frequent 3-sequences

< {1} {2} {3} >
< {1} {2 5} >
< {1} {5} {3} >
< {2} {3} {4} >
< {2 5} {3} >
< {3} {4} {5} >
< {5} {3 4} >

< {1} {2} {3} > : < {1} {2} {3} >, < {1} {2 5} >, < {1} {5} {3} >, < {2} {3} {4} >, < {2 5} {3} >, < {3} {4} {5} >, < {5} {3 4} >

< {1} {2 5} > : < {1} {2} {3} >, < {1} {2 5} >, < {1} {5} {3} >, < {2} {3} {4} >, < {2 5} {3} >, < {3} {4} {5} >, < {5} {3 4} >

< {1} {5} {3} > : < {1} {2} {3} >, < {1} {2 5} >, < {1} {5} {3} >, < {2} {3} {4} >, < {2 5} {3} >, < {3} {4} {5} >, < {5} {3 4} >

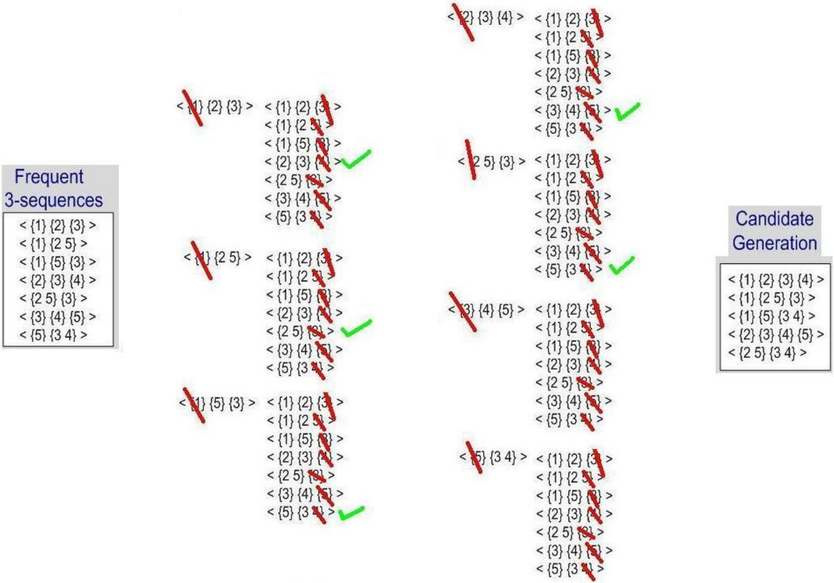
< {2} {3} {4} > : < {1} {2} {3} >, < {1} {2 5} >, < {1} {5} {3} >, < {2} {3} {4} >, < {2 5} {3} >, < {3} {4} {5} >, < {5} {3 4} >

< {2 5} {3} > : < {1} {2} {3} >, < {1} {2 5} >, < {1} {5} {3} >, < {2} {3} {4} >, < {2 5} {3} >, < {3} {4} {5} >, < {5} {3 4} >

< {3} {4} {5} > : < {1} {2} {3} >, < {1} {2 5} >, < {1} {5} {3} >, < {2} {3} {4} >, < {2 5} {3} >, < {3} {4} {5} >, < {5} {3 4} >

< {5} {3 4} > : < {1} {2} {3} >, < {1} {2 5} >, < {1} {5} {3} >, < {2} {3} {4} >, < {2 5} {3} >, < {3} {4} {5} >, < {5} {3 4} >

Candidate Generation



Frequent 3-sequences

< {1} {2} {3} >
< {1} {2 5} >
< {1} {5} {3} >
< {2} {3} {4} >
< {2 5} {3} >
< {3} {4} {5} >
< {5} {3 4} >

< ~~{1}~~ {2} {3} >

< {1} {2} ~~{3}~~ >
< {1} {2 ~~5}~~ >
< {1} {5} ~~{3}~~ >
< {2} {3} ~~{4}~~ > ✓
< {2 5} ~~{3}~~ >
< {3} {4} ~~{5}~~ >
< {5} {3 ~~4}~~ >

< ~~{1}~~ {2 5} >

< {1} {2} ~~{3}~~ >
< {1} {2 ~~5}~~ >
< {1} {5} ~~{3}~~ >
< {2} {3} ~~{4}~~ >
< {2 5} ~~{3}~~ > ✓
< {3} {4} ~~{5}~~ >
< {5} {3 ~~4}~~ >

< ~~{1}~~ {5} {3} >

< {1} {2} ~~{3}~~ >
< {1} {2 ~~5}~~ >
< {1} {5} ~~{3}~~ >
< {2} {3} ~~{4}~~ >
< {2 5} ~~{3}~~ >
< {3} {4} ~~{5}~~ >
< {5} {3 ~~4}~~ > ✓

< ~~{2}~~ {3} {4} >

< {1} {2} ~~{3}~~ >
< {1} {2 ~~5}~~ >
< {1} {5} ~~{3}~~ >
< {2} {3} ~~{4}~~ >
< {2 5} ~~{3}~~ >
< {3} {4} ~~{5}~~ > ✓
< {5} {3 ~~4}~~ >

< ~~{2~~ 5} {3} >

< {1} {2} ~~{3}~~ >
< {1} {2 ~~5}~~ >
< {1} {5} ~~{3}~~ >
< {2} {3} ~~{4}~~ >
< {2 5} ~~{3}~~ >
< {3} {4} ~~{5}~~ >
< {5} {3 ~~4}~~ > ✓

< ~~{3}~~ {4} {5} >

< {1} {2} ~~{3}~~ >
< {1} {2 ~~5}~~ >
< {1} {5} ~~{3}~~ >
< {2} {3} ~~{4}~~ >
< {2 5} ~~{3}~~ >
< {3} {4} ~~{5}~~ >
< {5} {3 ~~4}~~ >

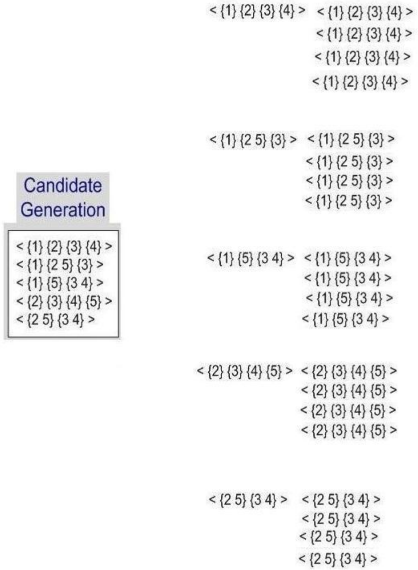
< ~~{5}~~ {3 4} >

< {1} {2} ~~{3}~~ >
< {1} {2 ~~5}~~ >
< {1} {5} ~~{3}~~ >
< {2} {3} ~~{4}~~ >
< {2 5} ~~{3}~~ >
< {3} {4} ~~{5}~~ >
< {5} {3 ~~4}~~ >

Candidate Generation

< {1} {2} {3} {4} >
< {1} {2 5} {3} >
< {1} {5} {3 4} >
< {2} {3} {4} {5} >
< {2 5} {3 4} >

Candidate Pruning



Candidate Generation

< {1} {2} {3} {4} >
< {1} {2 5} {3} >
< {1} {5} {3 4} >
< {2} {3} {4} {5} >
< {2 5} {3 4} >

< {1} {2} {3} {4} >
< {1} {2} {3} {4} >
< {1} {2} {3} {4} >
< {1} {2} {3} {4} >
< {1} {2} {3} {4} >

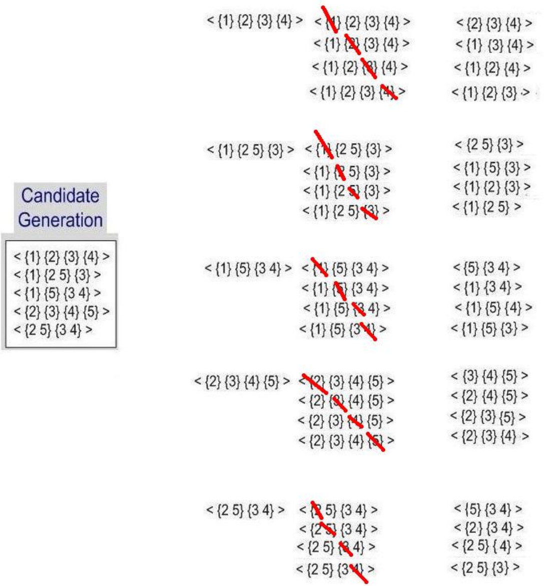
< {1} {2 5} {3} >
< {1} {2 5} {3} >
< {1} {2 5} {3} >
< {1} {2 5} {3} >
< {1} {2 5} {3} >

< {1} {5} {3 4} >
< {1} {5} {3 4} >
< {1} {5} {3 4} >
< {1} {5} {3 4} >
< {1} {5} {3 4} >

< {2} {3} {4} {5} >
< {2} {3} {4} {5} >
< {2} {3} {4} {5} >
< {2} {3} {4} {5} >
< {2} {3} {4} {5} >

< {2 5} {3 4} >
< {2 5} {3 4} >
< {2 5} {3 4} >
< {2 5} {3 4} >
< {2 5} {3 4} >

Candidate Pruning

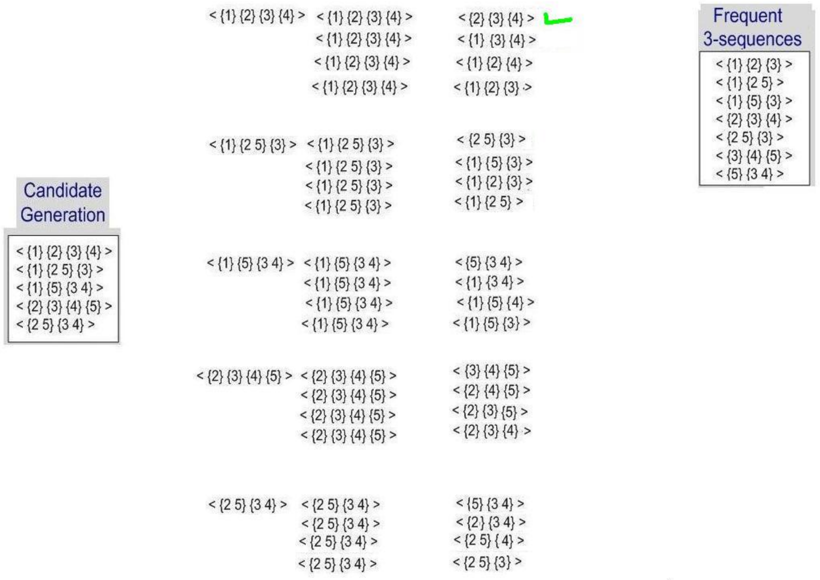


Candidate Generation

< {1} {2} {3} {4} >
< {1} {2 5} {3} >
< {1} {5} {3 4} >
< {2} {3} {4} {5} >
< {2 5} {3 4} >

< {1} {2} {3} {4} >	< {~~1~~} {2} {3} {4} >	< {2} {3} {4} >
	< {1} {~~2~~} {3} {4} >	< {1} {3} {4} >
	< {1} {2} {~~3~~} {4} >	< {1} {2} {4} >
	< {1} {2} {3} {~~4~~} >	< {1} {2} {3} >
< {1} {2 5} {3} >	< {~~1~~} {2 5} {3} >	< {2 5} {3} >
	< {1} {~~2~~ 5} {3} >	< {1} {5} {3} >
	< {1} {2 ~~5~~} {3} >	< {1} {2} {3} >
	< {1} {2 5} {~~3~~} >	< {1} {2 5} >
< {1} {5} {3 4} >	< {~~1~~} {5} {3 4} >	< {5} {3 4} >
	< {1} {~~5~~} {3 4} >	< {1} {3 4} >
	< {1} {5} {~~3~~ 4} >	< {1} {5} {4} >
	< {1} {5} {3 ~~4~~} >	< {1} {5} {3} >
< {2} {3} {4} {5} >	< {~~2~~} {3} {4} {5} >	< {3} {4} {5} >
	< {2} {~~3~~} {4} {5} >	< {2} {4} {5} >
	< {2} {3} {~~4~~} {5} >	< {2} {3} {5} >
	< {2} {3} {4} {~~5~~} >	< {2} {3} {4} >
< {2 5} {3 4} >	< {~~2~~ 5} {3 4} >	< {5} {3 4} >
	< {2 ~~5~~} {3 4} >	< {2} {3 4} >
	< {2 5} {~~3~~ 4} >	< {2 5} { 4} >
	< {2 5} {3 ~~4~~} >	< {2 5} {3} >

Candidate Pruning



Candidate Generation

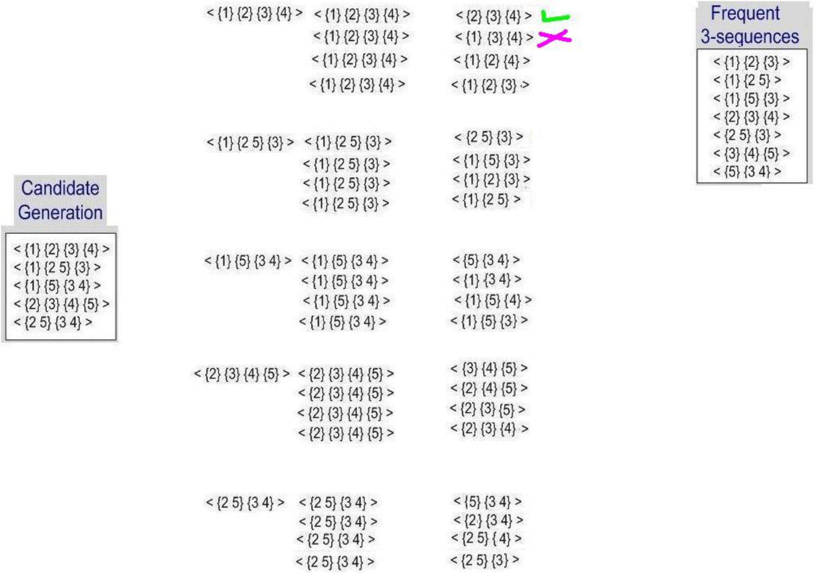
- < {1} {2} {3} {4} >
- < {1} {2 5} {3} >
- < {1} {5} {3 4} >
- < {2} {3} {4} {5} >
- < {2 5} {3 4} >

< {1} {2} {3} {4} >	< {1} {2} {3} {4} >	< {2} {3} {4} > ✓
	< {1} {2} {3} {4} >	< {1} {3} {4} >
	< {1} {2} {3} {4} >	< {1} {2} {4} >
	< {1} {2} {3} {4} >	< {1} {2} {3} >
< {1} {2 5} {3} >	< {1} {2 5} {3} >	< {2 5} {3} >
	< {1} {2 5} {3} >	< {1} {5} {3} >
	< {1} {2 5} {3} >	< {1} {2} {3} >
	< {1} {2 5} {3} >	< {1} {2 5} >
< {1} {5} {3 4} >	< {1} {5} {3 4} >	< {5} {3 4} >
	< {1} {5} {3 4} >	< {1} {3 4} >
	< {1} {5} {3 4} >	< {1} {5} {4} >
	< {1} {5} {3 4} >	< {1} {5} {3} >
< {2} {3} {4} {5} >	< {2} {3} {4} {5} >	< {3} {4} {5} >
	< {2} {3} {4} {5} >	< {2} {4} {5} >
	< {2} {3} {4} {5} >	< {2} {3} {5} >
	< {2} {3} {4} {5} >	< {2} {3} {4} >
< {2 5} {3 4} >	< {2 5} {3 4} >	< {5} {3 4} >
	< {2 5} {3 4} >	< {2} {3 4} >
	< {2 5} {3 4} >	< {2 5} { 4} >
	< {2 5} {3 4} >	< {2 5} {3} >

Frequent 3-sequences

- < {1} {2} {3} >
- < {1} {2 5} >
- < {1} {5} {3} >
- < {2} {3} {4} >
- < {2 5} {3} >
- < {3} {4} {5} >
- < {5} {3 4} >

Candidate Pruning



Candidate Generation

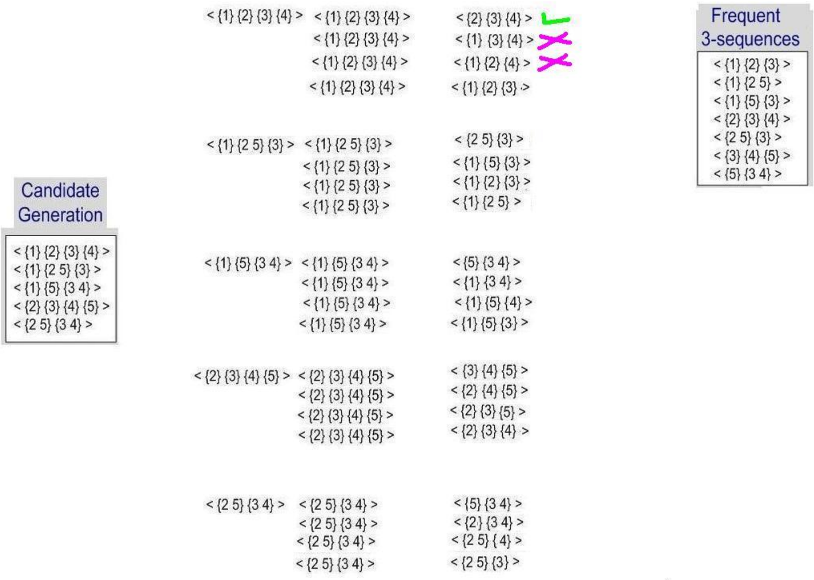
< {1} {2} {3} {4} >
< {1} {2 5} {3} >
< {1} {5} {3 4} >
< {2} {3} {4} {5} >
< {2 5} {3 4} >

< {1} {2} {3} {4} >	< {1} {2} {3} {4} >	< {2} {3} {4} > ✓
	< {1} {2} {3} {4} >	< {1} {3} {4} > ✗
	< {1} {2} {3} {4} >	< {1} {2} {4} >
	< {1} {2} {3} {4} >	< {1} {2} {3} >
< {1} {2 5} {3} >	< {1} {2 5} {3} >	< {2 5} {3} >
	< {1} {2 5} {3} >	< {1} {5} {3} >
	< {1} {2 5} {3} >	< {1} {2} {3} >
	< {1} {2 5} {3} >	< {1} {2 5} >
< {1} {5} {3 4} >	< {1} {5} {3 4} >	< {5} {3 4} >
	< {1} {5} {3 4} >	< {1} {3 4} >
	< {1} {5} {3 4} >	< {1} {5} {4} >
	< {1} {5} {3 4} >	< {1} {5} {3} >
< {2} {3} {4} {5} >	< {2} {3} {4} {5} >	< {3} {4} {5} >
	< {2} {3} {4} {5} >	< {2} {4} {5} >
	< {2} {3} {4} {5} >	< {2} {3} {5} >
	< {2} {3} {4} {5} >	< {2} {3} {4} >
< {2 5} {3 4} >	< {2 5} {3 4} >	< {5} {3 4} >
	< {2 5} {3 4} >	< {2} {3 4} >
	< {2 5} {3 4} >	< {2 5} { 4} >
	< {2 5} {3 4} >	< {2 5} {3} >

Frequent 3-sequences

< {1} {2} {3} >
< {1} {2 5} >
< {1} {5} {3} >
< {2} {3} {4} >
< {2 5} {3} >
< {3} {4} {5} >
< {5} {3 4} >

Candidate Pruning



Candidate Generation

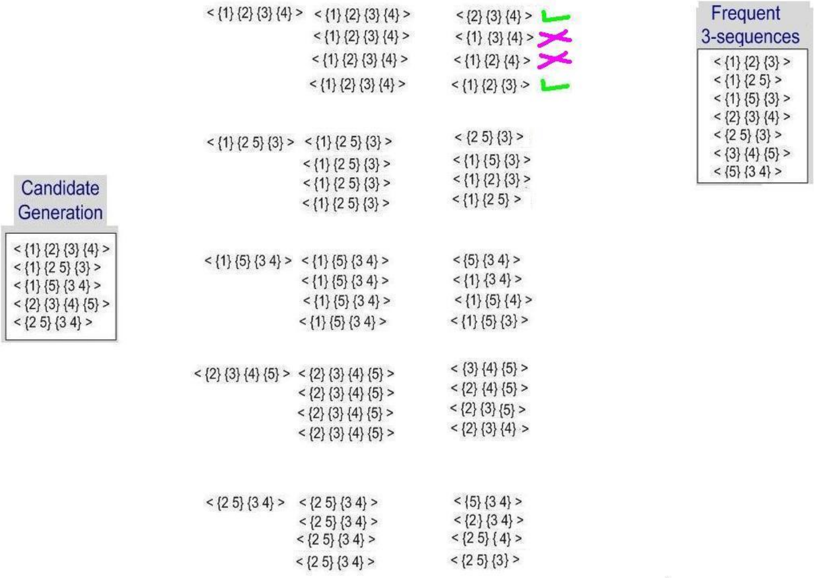
- < {1} {2} {3} {4} >
- < {1} {2 5} {3} >
- < {1} {5} {3 4} >
- < {2} {3} {4} {5} >
- < {2 5} {3 4} >

< {1} {2} {3} {4} >	< {1} {2} {3} {4} >	< {2} {3} {4} > ✓
	< {1} {2} {3} {4} >	< {1} {3} {4} > ✗
	< {1} {2} {3} {4} >	< {1} {2} {4} > ✗
	< {1} {2} {3} {4} >	< {1} {2} {3} >
< {1} {2 5} {3} >	< {1} {2 5} {3} >	< {2 5} {3} >
	< {1} {2 5} {3} >	< {1} {5} {3} >
	< {1} {2 5} {3} >	< {1} {2} {3} >
	< {1} {2 5} {3} >	< {1} {2 5} >
< {1} {5} {3 4} >	< {1} {5} {3 4} >	< {5} {3 4} >
	< {1} {5} {3 4} >	< {1} {3 4} >
	< {1} {5} {3 4} >	< {1} {5} {4} >
	< {1} {5} {3 4} >	< {1} {5} {3} >
< {2} {3} {4} {5} >	< {2} {3} {4} {5} >	< {3} {4} {5} >
	< {2} {3} {4} {5} >	< {2} {4} {5} >
	< {2} {3} {4} {5} >	< {2} {3} {5} >
	< {2} {3} {4} {5} >	< {2} {3} {4} >
< {2 5} {3 4} >	< {2 5} {3 4} >	< {5} {3 4} >
	< {2 5} {3 4} >	< {2} {3 4} >
	< {2 5} {3 4} >	< {2 5} { 4} >
	< {2 5} {3 4} >	< {2 5} {3} >

Frequent 3-sequences

- < {1} {2} {3} >
- < {1} {2 5} >
- < {1} {5} {3} >
- < {2} {3} {4} >
- < {2 5} {3} >
- < {3} {4} {5} >
- < {5} {3 4} >

Candidate Pruning



Candidate Generation

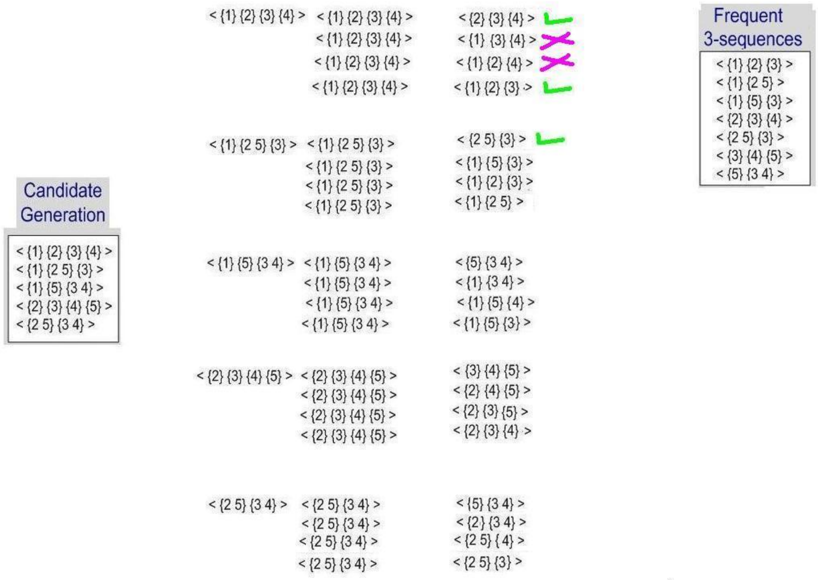
< {1} {2} {3} {4} >
< {1} {2 5} {3} >
< {1} {5} {3 4} >
< {2} {3} {4} {5} >
< {2 5} {3 4} >

< {1} {2} {3} {4} >	< {1} {2} {3} {4} >	< {2} {3} {4} > ✓
	< {1} {2} {3} {4} >	< {1} {3} {4} > ✗
	< {1} {2} {3} {4} >	< {1} {2} {4} > ✗
	< {1} {2} {3} {4} >	< {1} {2} {3} > ✓
< {1} {2 5} {3} >	< {1} {2 5} {3} >	< {2 5} {3} >
	< {1} {2 5} {3} >	< {1} {5} {3} >
	< {1} {2 5} {3} >	< {1} {2} {3} >
	< {1} {2 5} {3} >	< {1} {2 5} >
< {1} {5} {3 4} >	< {1} {5} {3 4} >	< {5} {3 4} >
	< {1} {5} {3 4} >	< {1} {3 4} >
	< {1} {5} {3 4} >	< {1} {5} {4} >
	< {1} {5} {3 4} >	< {1} {5} {3} >
< {2} {3} {4} {5} >	< {2} {3} {4} {5} >	< {3} {4} {5} >
	< {2} {3} {4} {5} >	< {2} {4} {5} >
	< {2} {3} {4} {5} >	< {2} {3} {5} >
	< {2} {3} {4} {5} >	< {2} {3} {4} >
< {2 5} {3 4} >	< {2 5} {3 4} >	< {5} {3 4} >
	< {2 5} {3 4} >	< {2} {3 4} >
	< {2 5} {3 4} >	< {2 5} { 4} >
	< {2 5} {3 4} >	< {2 5} {3} >

Frequent 3-sequences

< {1} {2} {3} >
< {1} {2 5} >
< {1} {5} {3} >
< {2} {3} {4} >
< {2 5} {3} >
< {3} {4} {5} >
< {5} {3 4} >

Candidate Pruning



Frequent 3-sequences

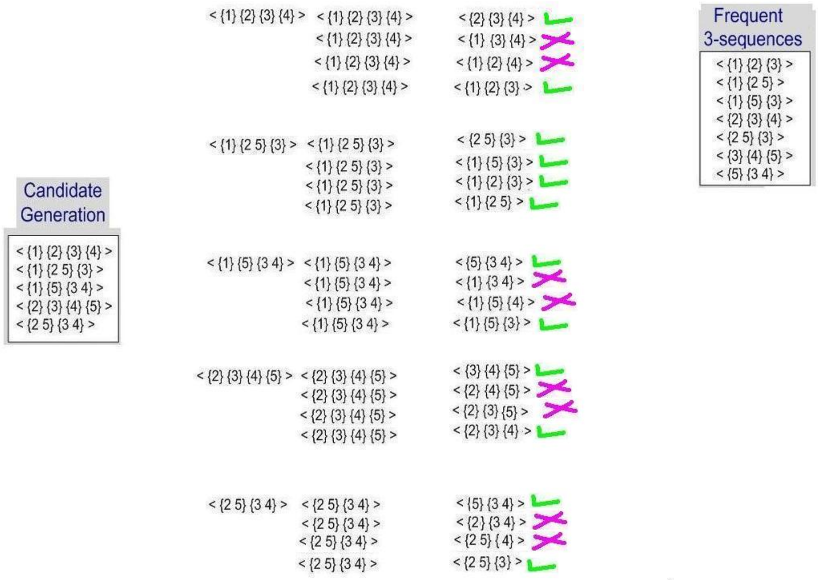
- < {1} {2} {3} >
- < {1} {2 5} >
- < {1} {5} {3} >
- < {2} {3} {4} >
- < {2 5} {3} >
- < {3} {4} {5} >
- < {5} {3 4} >

Candidate Generation

- < {1} {2} {3} {4} >
- < {1} {2 5} {3} >
- < {1} {5} {3 4} >
- < {2} {3} {4} {5} >
- < {2 5} {3 4} >

< {1} {2} {3} {4} >	< {1} {2} {3} {4} >	< {2} {3} {4} > ✓
	< {1} {2} {3} {4} >	< {1} {3} {4} > ✗
	< {1} {2} {3} {4} >	< {1} {2} {4} > ✗
	< {1} {2} {3} {4} >	< {1} {2} {3} > ✓
< {1} {2 5} {3} >	< {1} {2 5} {3} >	< {2 5} {3} > ✓
	< {1} {2 5} {3} >	< {1} {5} {3} >
	< {1} {2 5} {3} >	< {1} {2} {3} >
	< {1} {2 5} {3} >	< {1} {2 5} >
< {1} {5} {3 4} >	< {1} {5} {3 4} >	< {5} {3 4} >
	< {1} {5} {3 4} >	< {1} {3 4} >
	< {1} {5} {3 4} >	< {1} {5} {4} >
	< {1} {5} {3 4} >	< {1} {5} {3} >
< {2} {3} {4} {5} >	< {2} {3} {4} {5} >	< {3} {4} {5} >
	< {2} {3} {4} {5} >	< {2} {4} {5} >
	< {2} {3} {4} {5} >	< {2} {3} {5} >
	< {2} {3} {4} {5} >	< {2} {3} {4} >
< {2 5} {3 4} >	< {2 5} {3 4} >	< {5} {3 4} >
	< {2 5} {3 4} >	< {2} {3 4} >
	< {2 5} {3 4} >	< {2 5} { 4} >
	< {2 5} {3 4} >	< {2 5} {3} >

Candidate Pruning



Candidate Generation

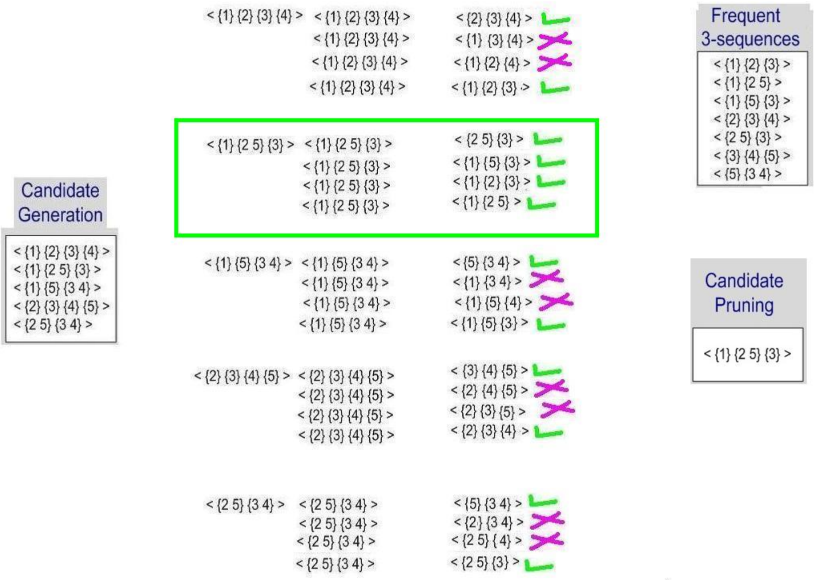
- < {1} {2} {3} {4} >
- < {1} {2 5} {3} >
- < {1} {5} {3 4} >
- < {2} {3} {4} {5} >
- < {2 5} {3 4} >

< {1} {2} {3} {4} >	< {1} {2} {3} {4} >	< {2} {3} {4} >	✓
	< {1} {2} {3} {4} >	< {1} {3} {4} >	✗
	< {1} {2} {3} {4} >	< {1} {2} {4} >	✗
	< {1} {2} {3} {4} >	< {1} {2} {3} >	✓
< {1} {2 5} {3} >	< {1} {2 5} {3} >	< {2 5} {3} >	✓
	< {1} {2 5} {3} >	< {1} {5} {3} >	✓
	< {1} {2 5} {3} >	< {1} {2} {3} >	✓
	< {1} {2 5} {3} >	< {1} {2 5} >	✓
< {1} {5} {3 4} >	< {1} {5} {3 4} >	< {5} {3 4} >	✓
	< {1} {5} {3 4} >	< {1} {3 4} >	✗
	< {1} {5} {3 4} >	< {1} {5} {4} >	✗
	< {1} {5} {3 4} >	< {1} {5} {3} >	✓
< {2} {3} {4} {5} >	< {2} {3} {4} {5} >	< {3} {4} {5} >	✓
	< {2} {3} {4} {5} >	< {2} {4} {5} >	✗
	< {2} {3} {4} {5} >	< {2} {3} {5} >	✗
	< {2} {3} {4} {5} >	< {2} {3} {4} >	✓
< {2 5} {3 4} >	< {2 5} {3 4} >	< {5} {3 4} >	✓
	< {2 5} {3 4} >	< {2} {3 4} >	✗
	< {2 5} {3 4} >	< {2 5} { 4} >	✗
	< {2 5} {3 4} >	< {2 5} {3} >	✓

Frequent 3-sequences

- < {1} {2} {3} >
- < {1} {2 5} >
- < {1} {5} {3} >
- < {2} {3} {4} >
- < {2 5} {3} >
- < {3} {4} {5} >
- < {5} {3 4} >

Candidate Pruning



Candidate Generation

- < {1} {2} {3} {4} >
- < {1} {2 5} {3} >
- < {1} {5} {3 4} >
- < {2} {3} {4} {5} >
- < {2 5} {3 4} >

< {1} {2} {3} {4} >	< {1} {2} {3} {4} >	< {2} {3} {4} >	✓
	< {1} {2} {3} {4} >	< {1} {3} {4} >	✗
	< {1} {2} {3} {4} >	< {1} {2} {4} >	✗
	< {1} {2} {3} {4} >	< {1} {2} {3} >	✓
< {1} {2 5} {3} >	< {1} {2 5} {3} >	< {2 5} {3} >	✓
	< {1} {2 5} {3} >	< {1} {5} {3} >	✓
	< {1} {2 5} {3} >	< {1} {2} {3} >	✓
	< {1} {2 5} {3} >	< {1} {2 5} >	✓
< {1} {5} {3 4} >	< {1} {5} {3 4} >	< {5} {3 4} >	✓
	< {1} {5} {3 4} >	< {1} {3 4} >	✗
	< {1} {5} {3 4} >	< {1} {5} {4} >	✗
	< {1} {5} {3 4} >	< {1} {5} {3} >	✓
< {2} {3} {4} {5} >	< {2} {3} {4} {5} >	< {3} {4} {5} >	✓
	< {2} {3} {4} {5} >	< {2} {4} {5} >	✗
	< {2} {3} {4} {5} >	< {2} {3} {5} >	✗
	< {2} {3} {4} {5} >	< {2} {3} {4} >	✓
< {2 5} {3 4} >	< {2 5} {3 4} >	< {5} {3 4} >	✓
	< {2 5} {3 4} >	< {2} {3 4} >	✗
	< {2 5} {3 4} >	< {2 5} { 4} >	✗
	< {2 5} {3 4} >	< {2 5} {3} >	✓

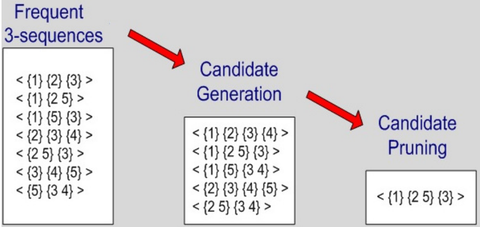
Frequent 3-sequences

- < {1} {2} {3} >
- < {1} {2 5} >
- < {1} {5} {3} >
- < {2} {3} {4} >
- < {2 5} {3} >
- < {3} {4} {5} >
- < {5} {3 4} >

Candidate Pruning

- < {1} {2 5} {3} >

GSP: Candidate Generation



Frequent 3-sequences

< {1} {2} {3} >
< {1} {2 5} >
< {1} {5} {3} >
< {2} {3} {4} >
< {2 5} {3} >
< {3} {4} {5} >
< {5} {3 4} >

Candidate Generation

< {1} {2} {3} {4} >
< {1} {2 5} {3} >
< {1} {5} {3 4} >
< {2} {3} {4} {5} >
< {2 5} {3 4} >

Candidate Pruning

< {1} {2 5} {3} >

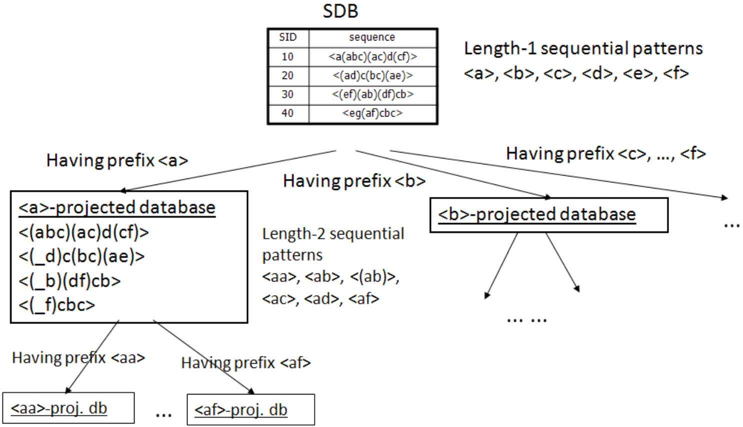
Issues:

- Huge number of candidate sets. *n* frequent 1-length candidate would generate $n^2 + \frac{n*(n-1)}{2}$ two-length candidate
- Multiple scans of the database
- Mining *n*-length sequential patterns need $\sum_{i=1}^{n} {}^{n}C_i = 2^n - 1$ number of short candidates. It is exponential

One can use prefix projections approach similar to FP-Growth

Pseudo-Projections[2]

When things can fit in main memory



SDB

SID	sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>

Length-1 sequential patterns
<a>, <b>, <c>, <d>, <e>, <f>

Having prefix <a>

Having prefix <b>

Having prefix <c>, ..., <f>

<a>-projected database
<(abc)(ac)d(cf)>
<(_d)c(bc)(ae)>
<(_b)(df)cb>
<(_f)cbc>

Length-2 sequential patterns
<aa>, <ab>, <(ab)>, <ac>, <ad>, <af>

<b>-projected database

...

... ...

Having prefix <aa>

Having prefix <af>

<aa>-proj. db

...

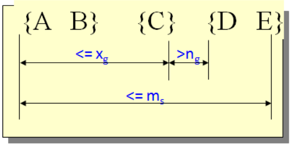
<af>-proj. db

[2] Han, Jiawei and Pei, Jian and Mortazavi-Asl, Behzad and Pinto, Helen and Chen, Qiming and Dayal, Umeshwar and Hsu, MC, “Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth” In proceedings of international conference on data engineering, pages 215–224, 2001

Application: Hotlink Assignment

- Hotlink are the mostly visited pages
- Website is modeled as graph, where pages are nodes and links are edges
- Web logs stores sequences of user clicks
- One sequence one session
- Intermediate pages could be navigational or target
- Use sequence mining to Mark start and end page of frequent sessions

Time Constraints

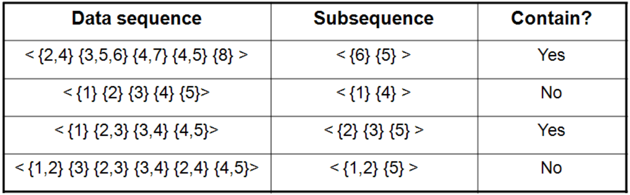


x_g: max-gap

n_g: min-gap

m_s: maximum span

$x_g = 2$, $n_g = 0$, $m_s = 4$



Data sequence	Subsequence	Contain?
< {2,4} {3,5,6} {4,7} {4,5} {8} >	< {6} {5} >	Yes
< {1} {2} {3} {4} {5}>	< {1} {4} >	No
< {1} {2,3} {3,4} {4,5}>	< {2} {3} {5} >	Yes
< {1,2} {3} {2,3} {3,4} {2,4} {4,5}>	< {1,2} {5} >	No

Approaches

1. Mine without timing constraint and post-process discovered patterns
2. Modify GSP to directly prune candidates violating timing

Anti Monotone Property

If a set is frequent all its subset must be frequent. If a set fails the test all its super set also must fail.

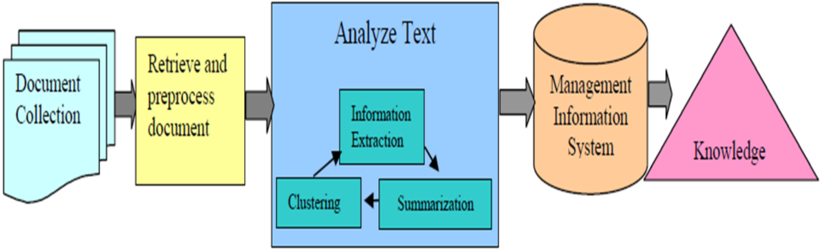
Consider sequences $A =< \{1, 2, 4\}, \{2, 3\}, \{5\} >$
$B =< \{1, 2\}, \{2, 3, 4\} >$ $C =< \{1, 2\}, \{2, 3, 4\}, \{2, 4, 5\} >$
$D =< \{2\}, \{3, 4\}, \{4, 5\} >$ $E =< \{1, 3\}, \{2, 4, 5\} >$

- Let $x_g = 1(max - gap)$, $n_g = 0(min - gap)$, $m_s = 5(maxspan)$, $minSup = 60\%$
- What is support for $< \{2\}, \{5\} >$? 40%
- What is support for $< \{2\}, \{3\}, \{5\} >$? 60%

Anti Monotone Property does **not** holds, so these properties can not be pushed in GSP

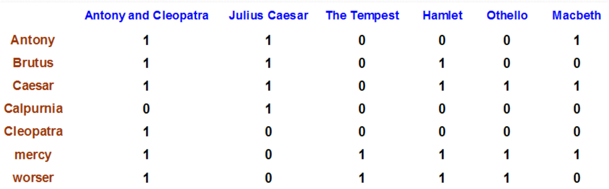
Text Mining

- Computer are bad to handle slang, spelling variations, contextual meaning and unstructured data
- Text is less structured
- Applications involves 1) Information Extraction, 2) Topic Tracking, 3) Summarization, 4) Categorization, 5) Clustering, 6) Concept Linkage, 7) Information Visualization, 8) Question Answering, *etc.*
- Starts with 1) Identity keywords and phrases, 2) Relationship within text



Text Representation

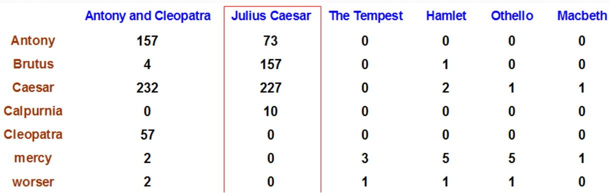
- Binary term-document incidence matrix



	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Document is represented as a binary vector $\in \{0, 1\}^{|V|}$

- Term-document count matrices



	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

Document is represented as a count vector $\in N^{|V|}$

Text Representation

- Bag of words: order of words is not important. See “Jon is lighter than Bob” and “Bob is lighter than Jon”
- Term frequency ($tf_{t,d}$): is number of times the term t occurs in document d. Relevance may not increase proportionally with term frequency.
- Log-frequency weighting: $w_{t,d} = \log(tf_{t,d})$ if $tf_{t,d} > 0$ otherwise zero. Consider following matching score b/w two documents

$$score = \sum_{t \in D_1 \cap D_2} w_{t,D_1}$$

Score is zero if no term of query document D_2 is present in D_1.

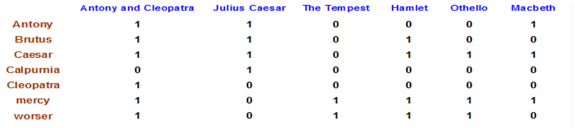
Text Representation

- **Document frequency:** Frequent terms are less informative than rare terms. df_t is number of documents that contain term t.
- **Inverse document frequency:** If we have N documents then $idf_t = \log(N/df_t)$
- **Collection frequency:** How many times term t appeared in all the document.
- **tfidf weighting:** $tfidf_{t,d} = \log(1 + tf_{t,d}) \times \log(N/df_t)$

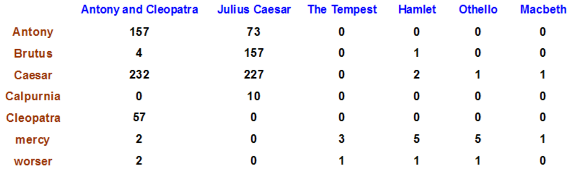
$$score = \sum_{t \in q \cap d} tfidf_{t,d}$$

This is the most used method to determine similarity.

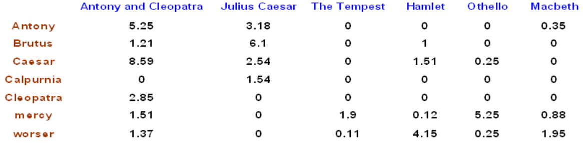
Text Representation



	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0



	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0



	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	5.25	3.18	0	0	0	0.35
Brutus	1.21	6.1	0	1	0	0
Caesar	8.59	2.54	0	1.51	0.25	0
Calpurnia	0	1.54	0	0	0	0
Cleopatra	2.85	0	0	0	0	0
mercy	1.51	0	1.9	0.12	5.25	0.88
worser	1.37	0	0.11	4.15	0.25	1.95

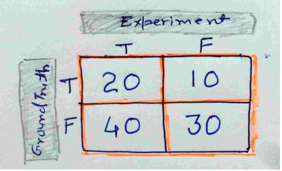
- Generally dimensionality reduction is also required
- How document classification? use k-NN, Naive Bayes, SVM ...
- How document clustering? use k-Means, Hierarchical, Agglomerative

Statistics

There were 100 images in a box. 30 of them were containing lion. I asked Bob to separate all the pics of lion. He showed me 60 but, lion was not in 40 of them.

- True positives (TP): 20
- True negatives (TN): 30
- T1-Error: False positives (FP): 40
- T2-Error: False negatives (FN): 10

Confusion Matrix



		Experiment	
		T	F
Ground Truth	T	20	10
	F	40	30

Accuracy: ((20+30)/100)*100%,
Precision: (20/60)*100%,
Recall (true positive rate or Sensitivity): (20/(20+10))*100%,
Specificity (true negative rate): (30/(40+30))*100%,
F Score: (Precision+Recall)/2,
F1 Measure: Harmonic mean of Precision and Recall

Thank You!

Thank you very much for your attention!

Queries ?