AUTHORSHIP CLASSIFICATION: A SYNTACTIC TREE MINING APPROACH

SANGKYUM KIM, HYUNGSUL KIM, TIM WENINGER, JIAWEI HAN

DEPT OF COMPUTER SCIENCE
UNIV OF ILLINOIS AT URBANA-CHAMPAIGN

UP@KDD'10 (July 2010, Washington DC)

Outline

- Background
- k-ee subtree pattern
- Two-step Discriminative Pattern Mining
- Experiments
- Conclusions

Document Clustering/Classification

- Topic vs. Genre vs. Authorship

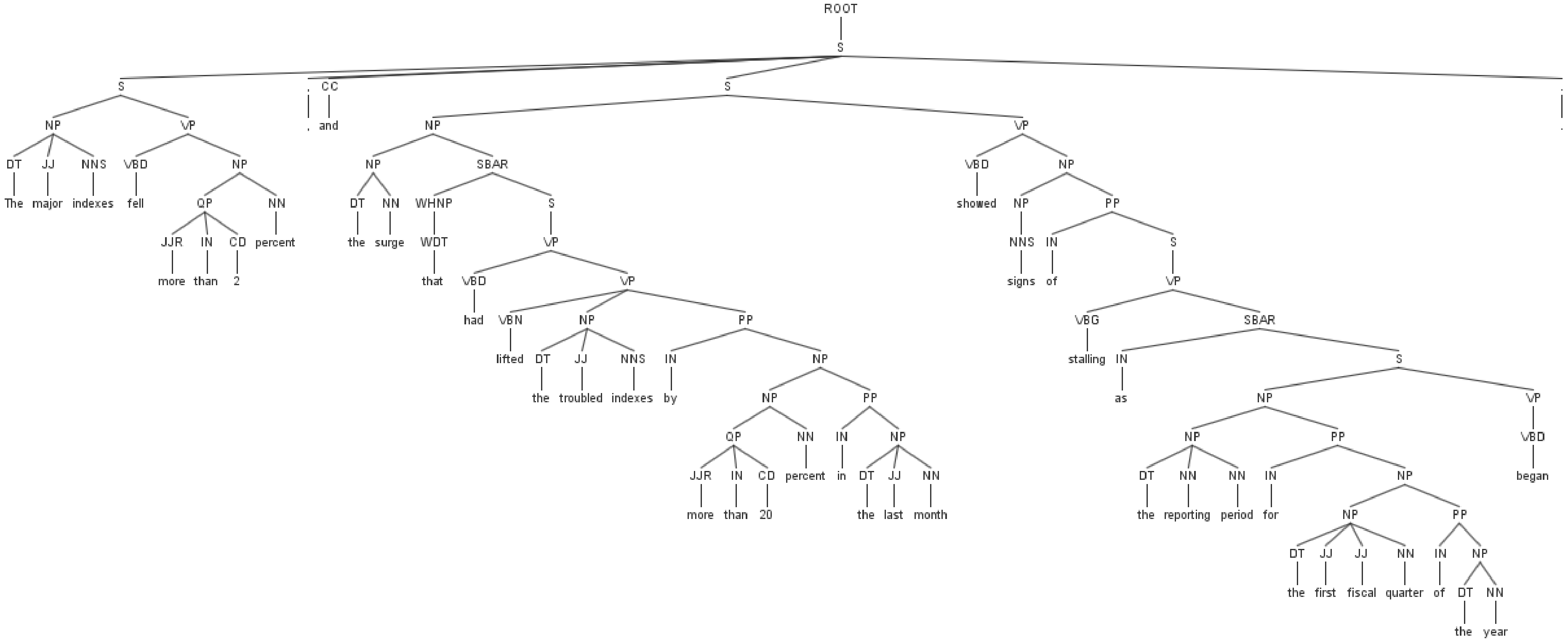
Topic	Genre	Authorship
Kennedy	John F. Kennedy	John F. Kennedy/News
• John F. Kennedy • JFK airport • Kennedy space center	• Blog • News article • Movie review • Academic report	• Writer 1 • Writer 2
• Subject-oriented words	• Punctuation marks • Simple common words • Genre specific words	• Function/syntactic words • POS tags • Rewrite rules

Existing Features for Authorship Classification

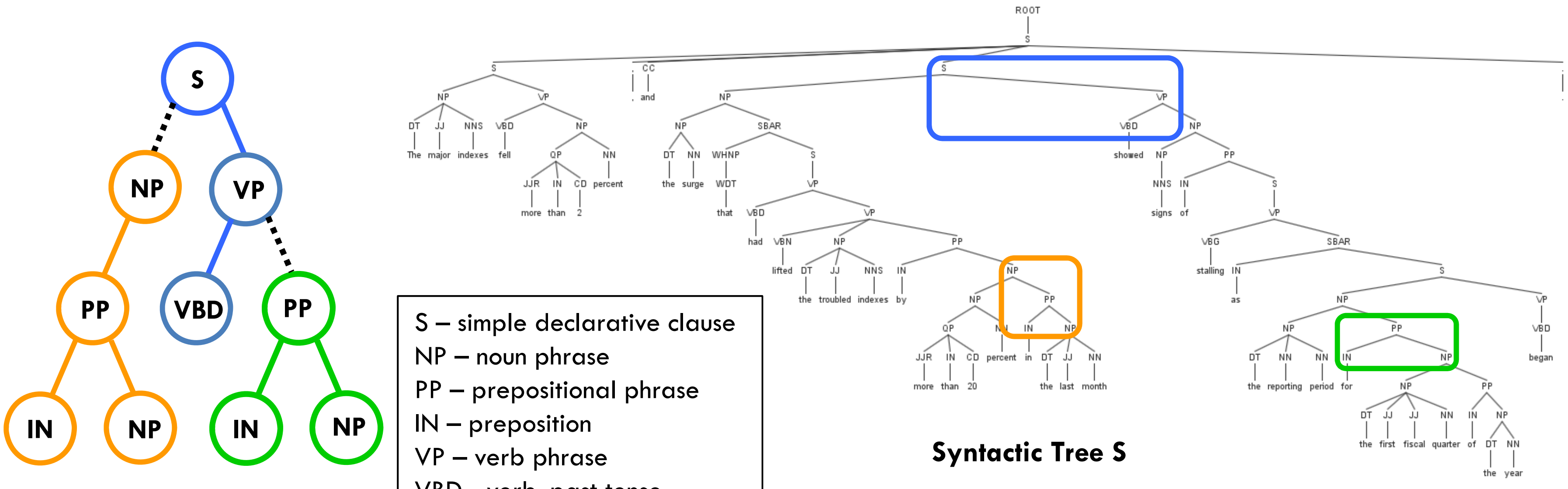
- Existing features for authorship classification
  - Function Words
    - the most common words (the, and, of, that, …)
    - little semantic content of their own but usually indicate a grammatical relationship or generic property
  - Part-Of-Speech tags
    - verb, noun, pronoun, adjective, adverb, preposition, ...
    - explains not what the word *is*, but how the word *is used*.
  - Rewrite Rules
    - $X \rightarrow Y_1 + Y_2 + \ldots + Y_n$
    - e.g. NP → DT+ JJ + JJ + NN

Syntactic Tree

Example. The major indexes fell more than 2 percent, and the surge that had lifted the troubled indexes by more than 20 percent in the last month showed signs of stalling as the reporting period for the first fiscal quarter of the year began.


ROOT
S
S , CC S .
, and .
NP VP NP VP
DT JJ NNS VBD NP NP SBAR VBD NP
The major indexes fell QP NN DT NN WHNP S showed NP PP
JJR IN CD percent the surge WDT VP NNS IN S
more than 2 that VBD VP signs of VP
had VBN NP PP VBG SBAR
lifted DT JJ NNS IN NP stalling IN S
the troubled indexes by NP PP as NP VP
QP NN IN NP NP PP VBD
JJR IN CD percent in DT JJ NN DT NN NN IN NP began
more than 20 the last month the reporting period for NP PP
DT JJ JJ NN IN NP
the first fiscal quarter of DT NN
the year

k-embedded-edge Subtree Pattern



S – simple declarative clause
NP – noun phrase
PP – prepositional phrase
IN – preposition
VP – verb phrase
VBD - verb, past tense

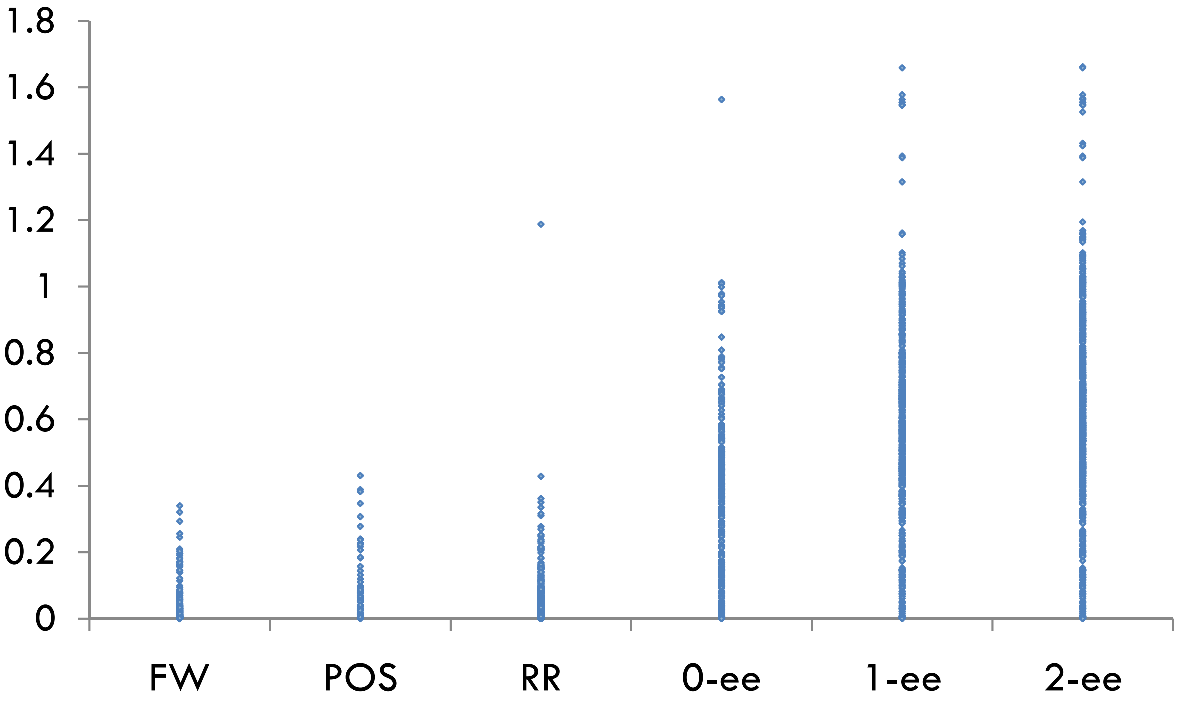
Pattern t

Syntactic Tree S

Example. The major indexes fell more than 2 percent, and the surge that had lifted the troubled indexes by more than 20 percent *in the last month* *showed* signs of stalling as the reporting period *for the first fiscal quarter of the year* began.

A 2-ee subtree pattern t is mined from two *NY Times* journalists Jack Healy and Eric Dash who worked in the same business department. On average, 21.2% of Jack's sentences contained t while only 7.2% of Eric's sentences contained t.}

Discriminative Score (Fisher Score)

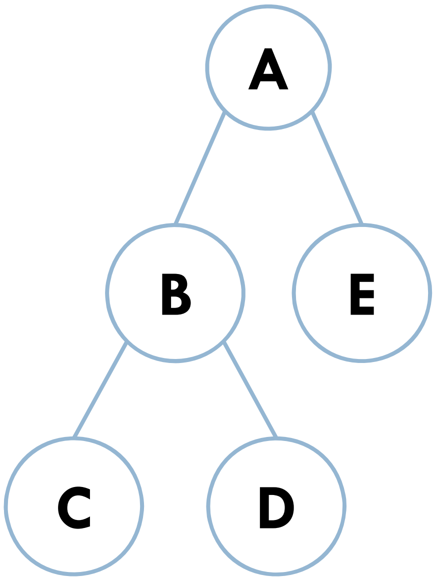


Two-Step Discriminative Pattern Mining

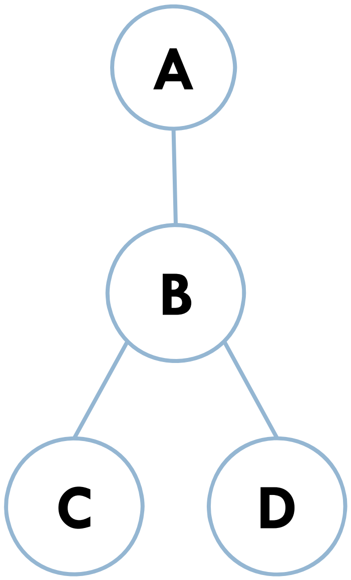
- <Step 1>
  - Mine closed frequent k-ee subtree patterns
    - Pattern-growth approach
  - Pruning with
    - Minimum support
    - Closed checking (backward/forward extension pruning)

- <Step 2>
  - Select discriminative patterns

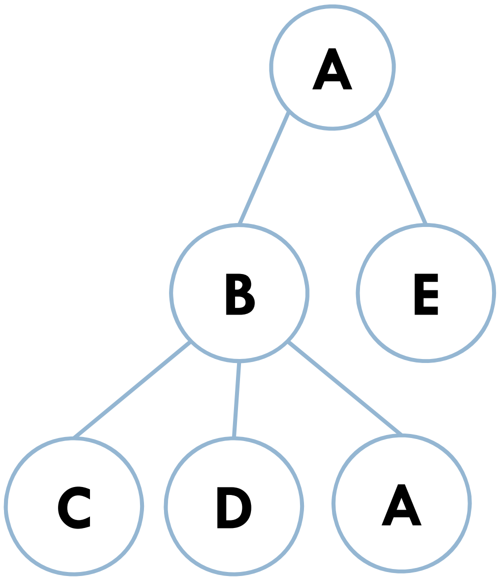
Toy DB



Sentence S_1

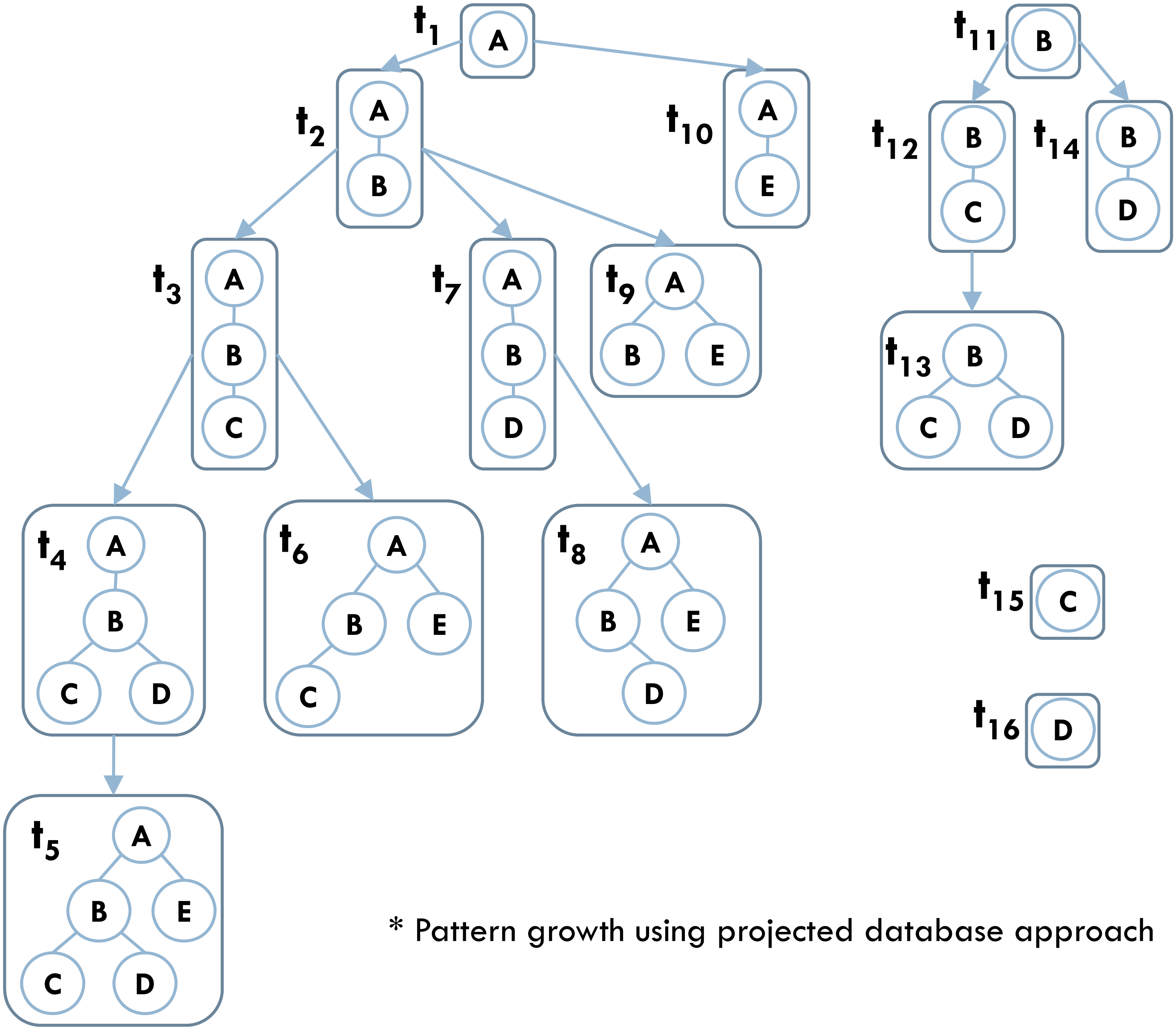


Sentence S_2



Sentence S_3

minsup=2/3

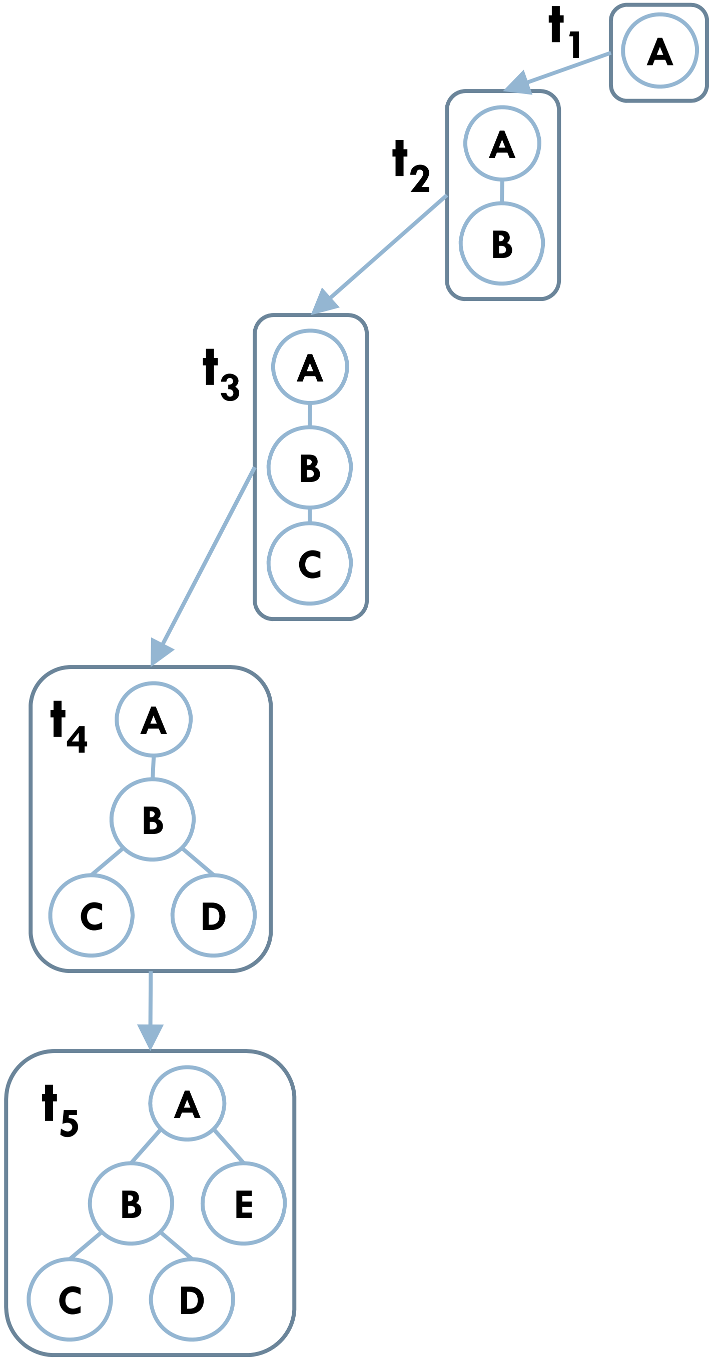


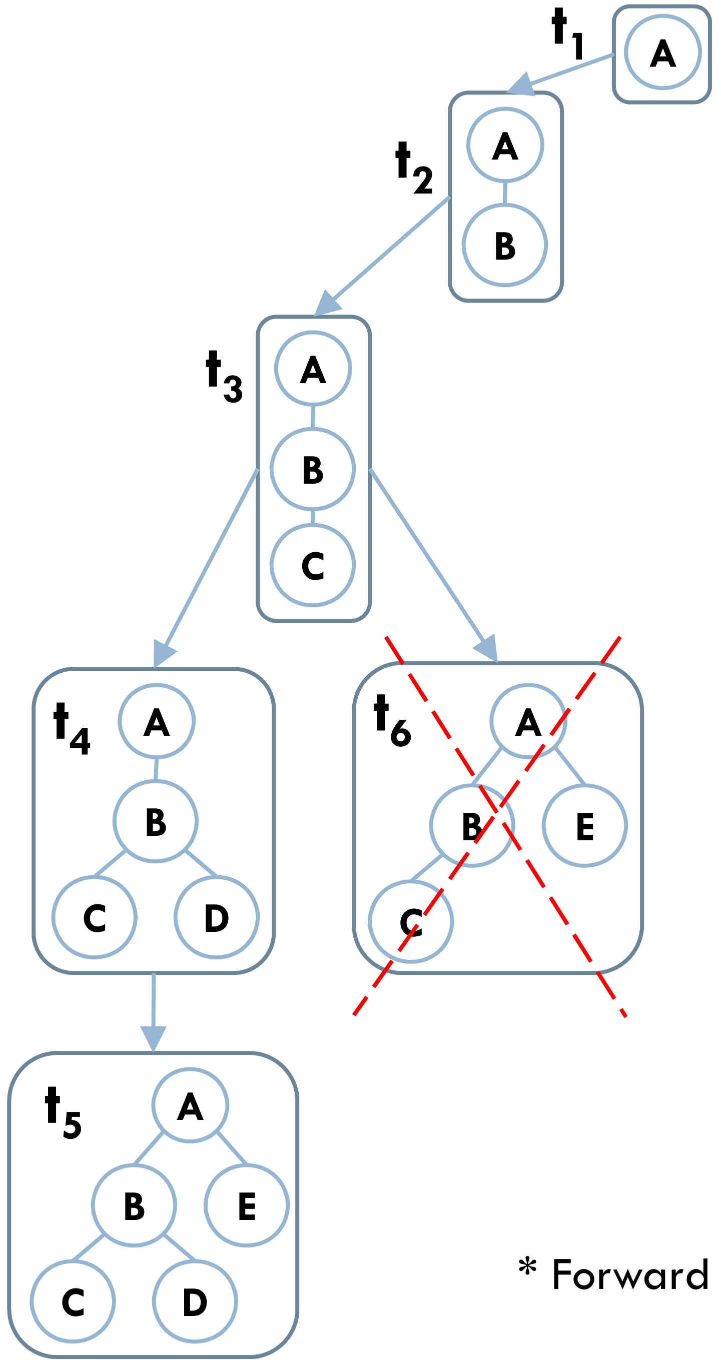
* Pattern growth using projected database approach

Rules for Pruning

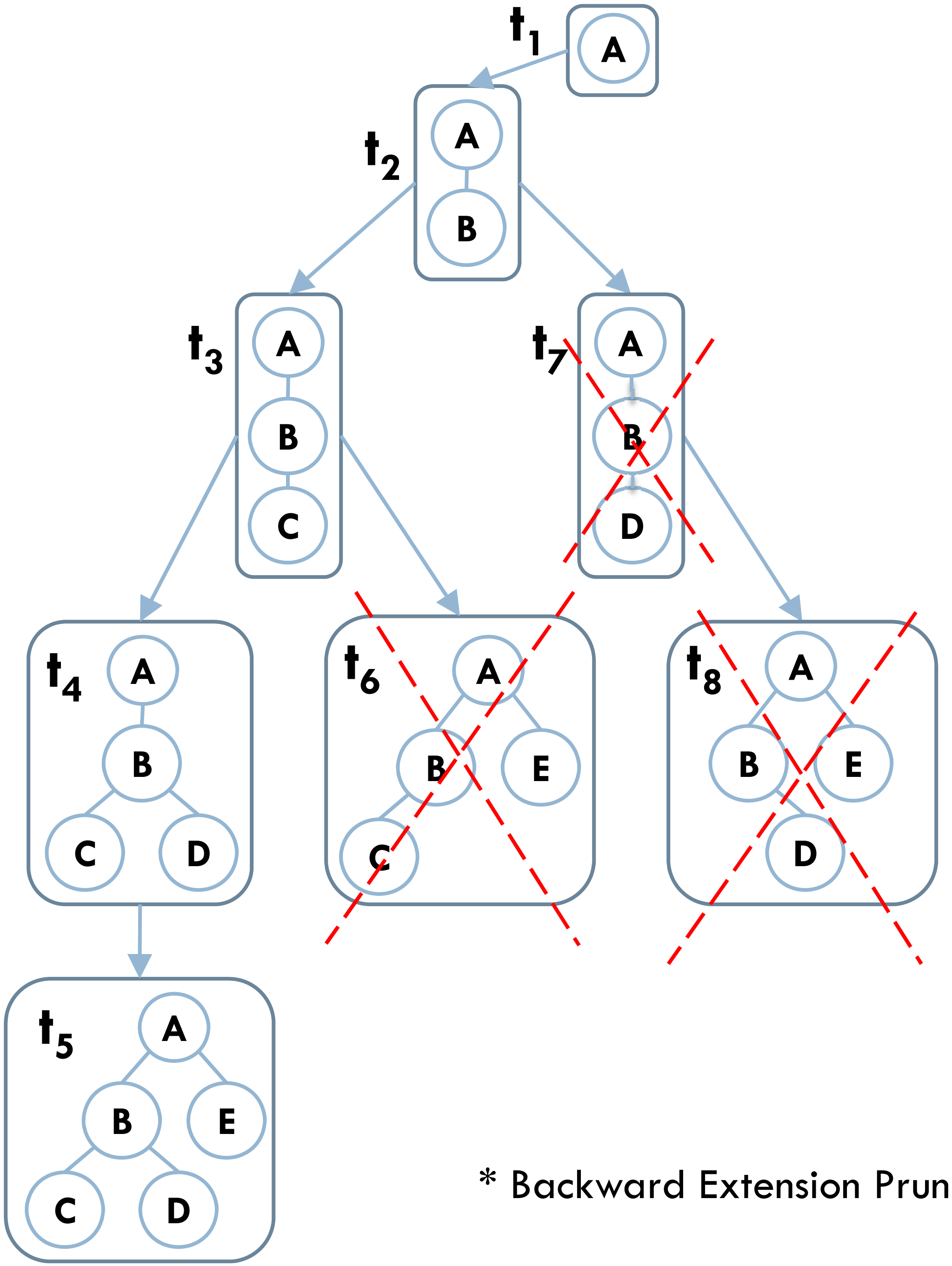
- Backward Extension Pruning
  - If there exists a backward extension node for a tree pattern t, then we do not need to extend t.
- Forward Extension Pruning
  - If there exists a forward extension node at node v in t, then we do not need to extend t by adding new rightmost nodes to any proper ancestor of v.

Adapted from CMTREEMINER (TKDE'05)

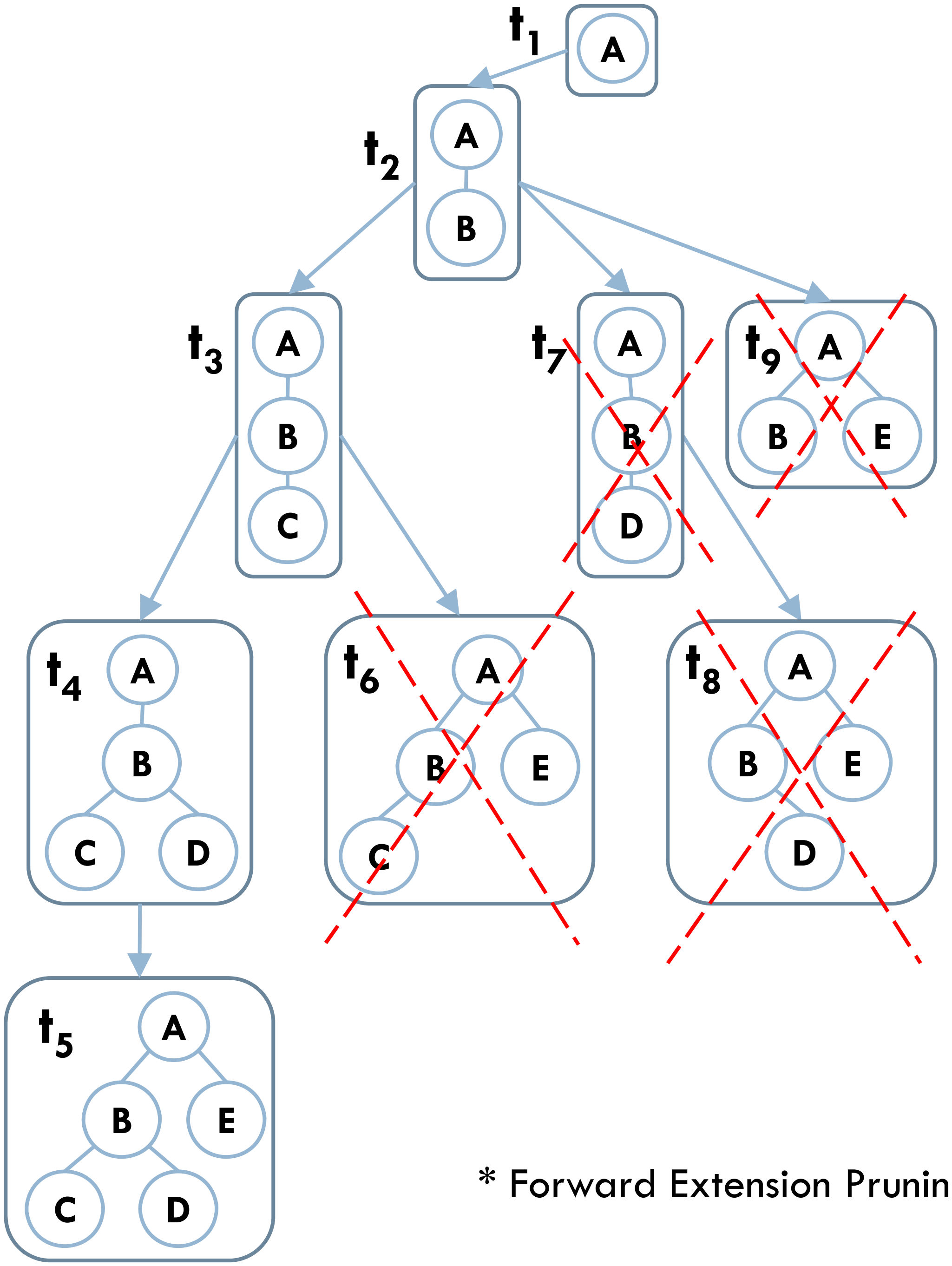

t_1
A
t_2
A
B
t_3
A
B
C
t_4
A
B
C
D
t_5
A
B
E
C
D



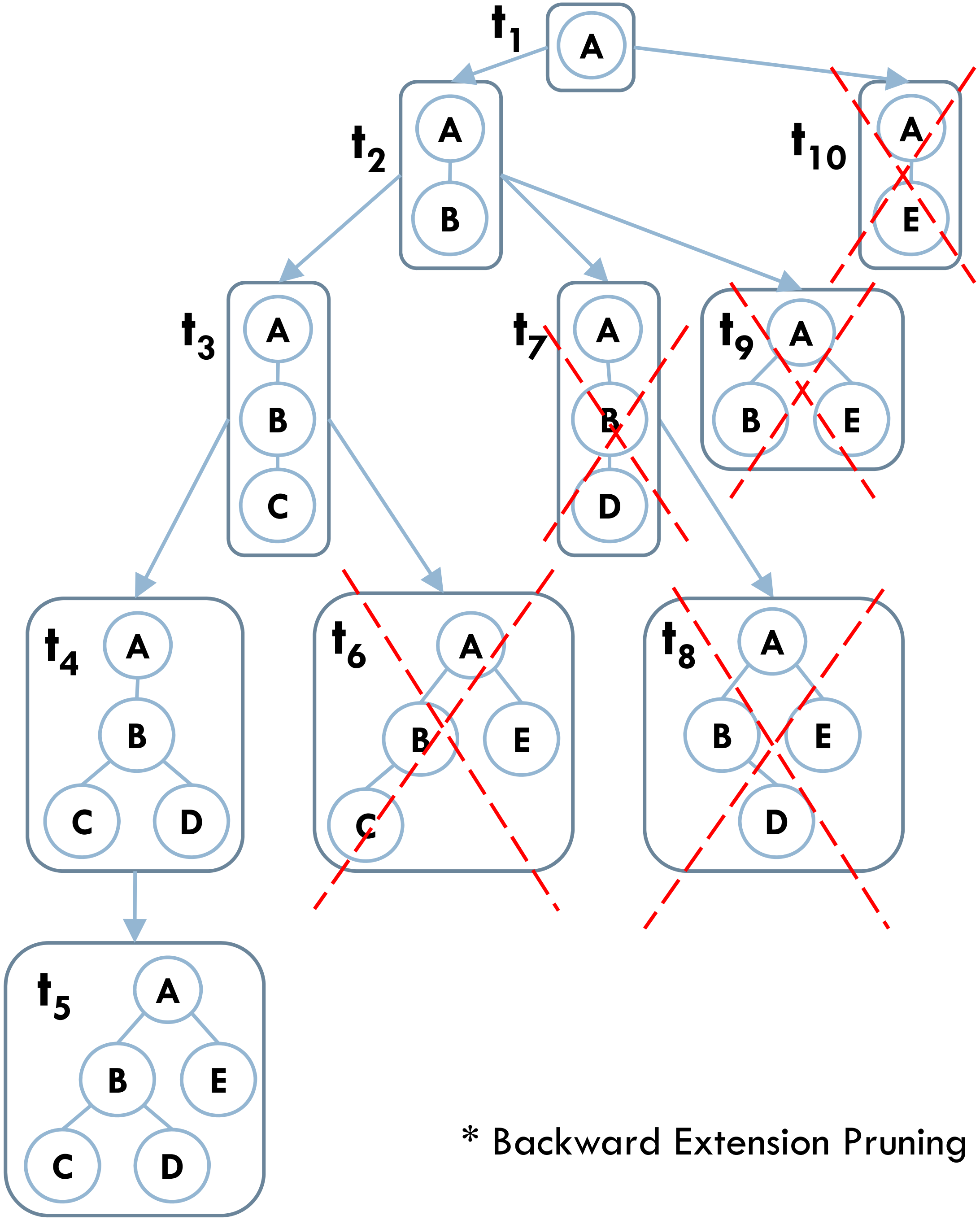
* Forward Extension Pruning



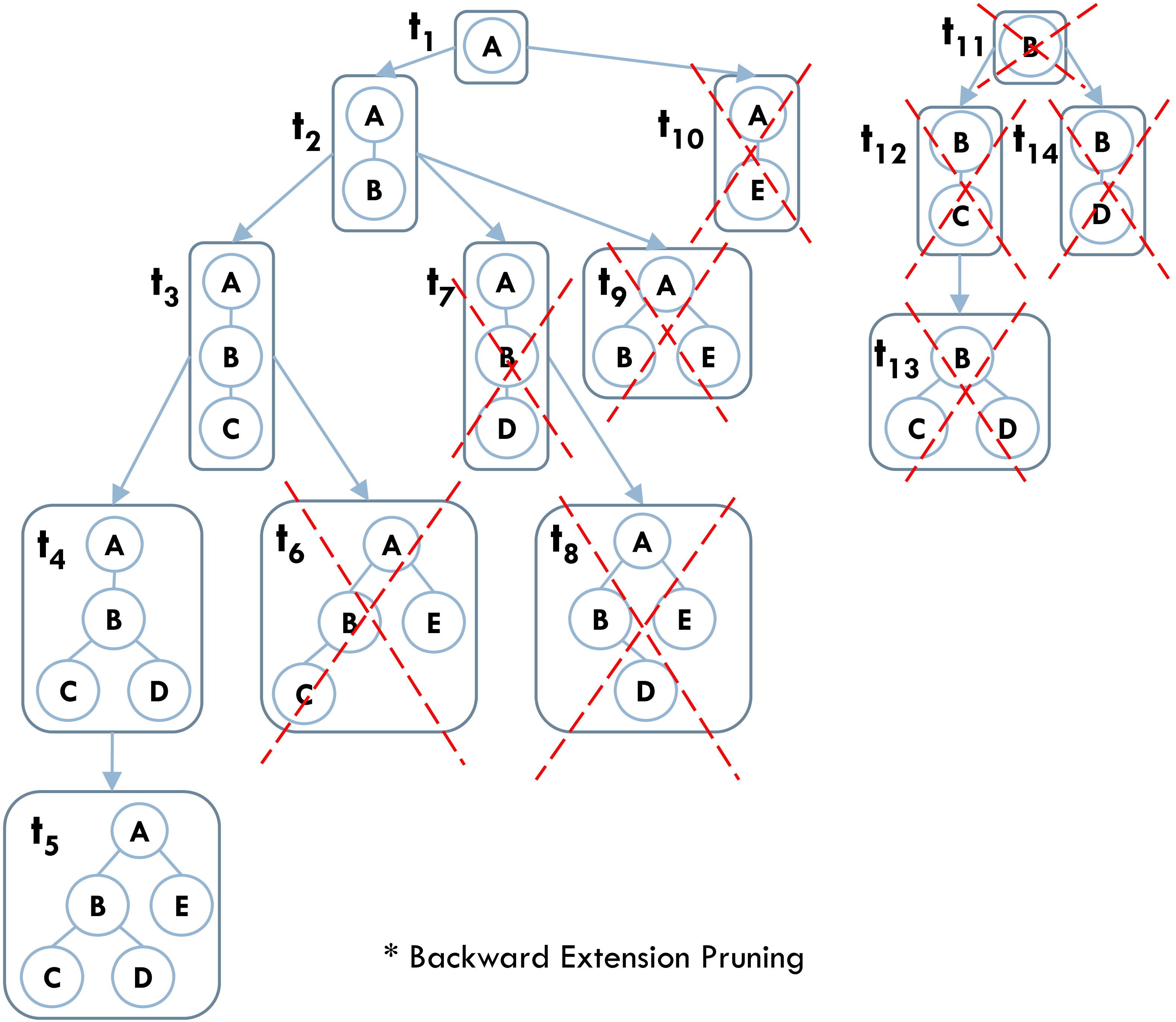
* Backward Extension Pruning



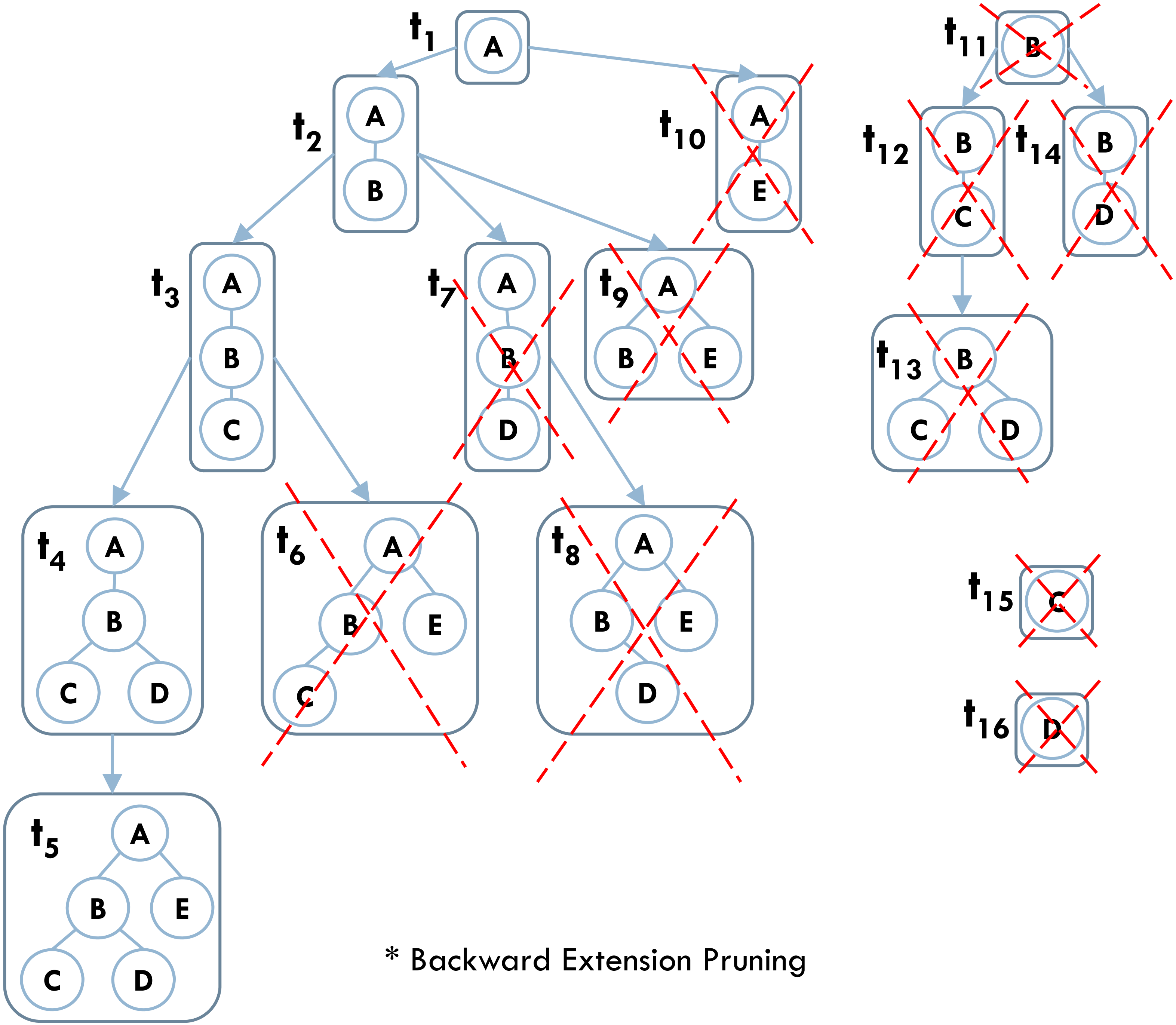
* Forward Extension Pruning



* Backward Extension Pruning



* Backward Extension Pruning



* Backward Extension Pruning

Experiments

- □ Data Sets (from *NYTimes.com*)

	# Authors	# Docs	# Sentences	# Words
News Articles	4	400	19K	381K
Movie Reviews	4	2K	51K	1.3M

- □ Size of Comparison Feature Sets

	FW	POS	RR	0-ee	1-ee	2-ee
News Articles	308	70	4K	280	560	790
Movie Reviews	308	70	9K	560	1.3K	2K

Experiments

- Accuracy (News Articles)

	FW	POS	RR	0-ee	1-ee	2-ee
N12	91.5	87	94	**96**	95	95.5
N13	94	85	91	97.5	**98**	97.5
N14	95.5	92.5	96	94.5	**96.5**	95
N23	95	92.5	92.5	96.5	98.5	**99**
N24	97	95.5	97.5	**98.5**	**98.5**	**98.5**
N34	80.5	67.5	67.5	88.5	**90**	**90**
AVG	92.3	86.7	89.8	95.3	**96.1**	96

Experiments

- □ Accuracy (Movie Reviews)

	FW	POS	RR	0-ee	1-ee	2-ee
N12	92.8	81	88	92.48	**94.26**	94.22
N13	93.6	92.5	92.7	95.22	95.06	**95.8**
N14	92.1	88	94.2	97	97.4	**97.7**
N23	94.4	92.8	94.8	97.58	**97.92**	97.58
N24	93.1	91	92.9	95.22	96.04	**96.32**
N34	93.1	88.6	94.9	97.12	**97.22**	97.12
AVG	93.2	89	92.9	95.8	96.3	**96.5**

Conclusions

- k-ee subtree pattern
  - Contains rich meaningful syntactic information
  - Bottleneck: Too many patterns
    - Adapt pruning methods of frequent and closed pattern mining
    - Mine only discriminative patterns
- Future work
  - Direct discriminative pattern mining
    - Two-step mining approach is still expensive
    - Avoid previous approaches of iteratively mining top-1 discriminative pattern