Thesis Meeting 11

kense, for the thesis

Review

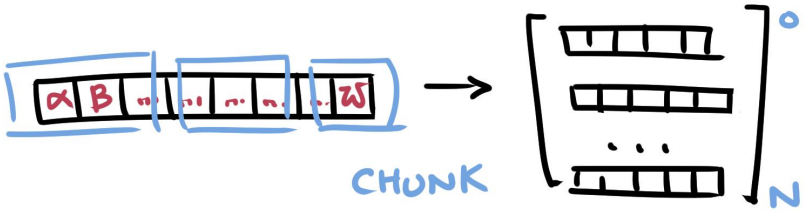
1. Develop a way to mine more patterns to integrate into weighting
2. Make more specific the labels used in the labelset.

PrefixSpan

- Sequential pattern mining
- Frequency based analysis
- Is solving a specific problem of pattern detection
- Terms: Sequence, element, item, sub-sequence, projected database, prefix, postfix, scanning, sequential pattern
- The important principle: “any super-pattern of a non-frequent pattern cannot be frequent”

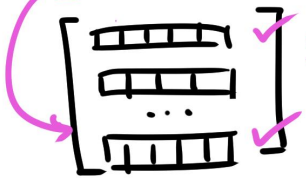
PrefixSpan Variation

- We’re looking for patterns isolated in each trace rather than patterns that emerge as common between all traces.
- Important principle from before still holds, however.
- To create a database, we chunk the trace into arbitrary parts. Then perform the algorithm as intended.



① Create the "bucket"

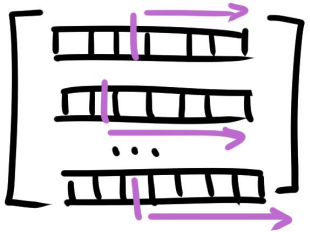
{α: 3, B: 2, ..., W: 1}



Search for existence, $0 \leq Val \leq N$

Drop entries whose Val < min_supp

② create "Projections"

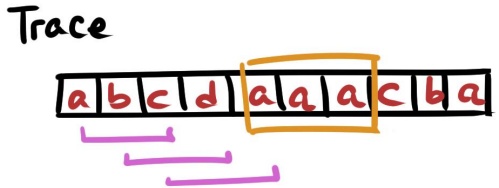


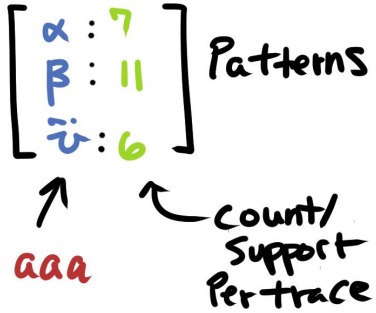
α

for all Patterns in the bucket.

Integration to Weighting

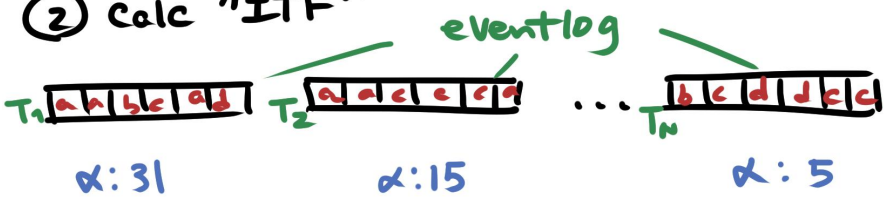
- Follows the principles of TF-IDF.
- Terms are replaced with patterns, and documents are replaced with traces.
- PF-ITF (Pattern Frequency Inverse Trace Frequency).
- Two-step Process:
  - Use PF-ITF to determine pattern importance (form a ranking).
  - Use ranking as a contribution metric to modify the weight of each action.





① Search for pattern "PF"

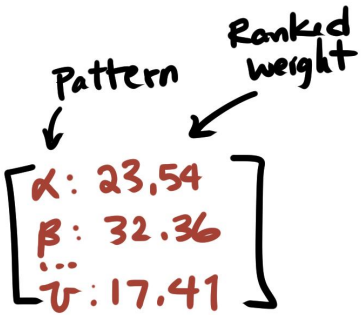
② Calc "ITF"



$$ITF = \log\left(\frac{\#\text{Patterns}}{PF_\alpha}\right)$$

③ Combine for entire eventlog

Ranked Patterns = PF * ITF



```
{'closed.question, give.statement': 32.94951737460809,
 'give.statement, give.statement, closed.question': 31.704990684331612,
 'x, x': 28.169095232978965,
 'recall, misc': 27.858926878661443,
 'use.social.convetion, use.social.convetion, use.social.convetion': 26.764067806412534,
 'give.statement, give.statement, give.statement, give.opinion': 25.74000594017879,
 'misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc': 24.978919395898856,
 'misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc': 24.978919395898856,
 'x, relax.atmosphere': 24.216437957464258,
 'give.statement, closed.question': 23.653370759545645,
 'misc, misc, misc, misc, misc, misc, misc, misc, misc': 23.325630060826683,
 'misc, misc, misc, misc, misc, misc, misc, misc, misc, misc': 23.325630060826683,
 'misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc': 23.325630060826683,
 'misc, misc, misc, misc, misc, misc, misc, misc': 22.894558015522527,
 'misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc': 22.594078081016285,
 'misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc': 22.594078081016285,
 'misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc, misc': 22.594078081016285,
 'use.social.convetion, use.social.convetion': 21.799143289520085,
 'give.statement, give.statement, give.statement, give.statement, give.opinion': 21.799143289520085,
 'misc, x': 21.799143289520085
```

Changes to Weighting Similarity

- Original weighting-modification to edit-distance doesn’t make sense once in new context of adding specific patterns.
- Augmented weights (previous + new weights) can be compared using cosine similarity.
- To compensate for differing lengths, we pad the shorter trace to maintain the distribution of the weights, thus penalizing high length differences.

Results (1/2)

- Trace similarity shows good promise, traces from the same show are showing ~0.90, while the outlier is around ~0.30.
- Expect this trend of values to be good for separation in clustering.
- Still have to integrate these values to our other metrics.

Label Expansion Considerations

- Noticed that these labels weren’t covered:
  - exclamation (previously under relax.atmosphere or give.statement, can cover generic exclamations: “Woo!” and commands, “Off you go!”)
  - inner.dialogue (not something present in live interview shows, but present in documentary and other media, separated from recall, which has connotation of the past)
  - ???
- Still have to modify and integrate.

//Paper Writing Segment

Please refer to the Google Docs page, which will be screen-shared, editing can be done live and more extensive changes can be noted.