Mining Closed Sequential Patterns in Large Datasets

Presenter: Ildar Nurgaliev
Lab: Dainfos

Main idea

Instead of mining the complete set of frequent subsequences we mine frequent *closed subsequences*

Benefits

- can mine really long sequences
- produce significantly less number of discovered frequent sequences

Preliminary Concepts

Sequence

- items: $I = \{i_1, i_2, ..., i_m\}$
- itemset (t_i): $t_i \subseteq I$
- sequence (ordered list): $s = \langle t_1, t_2, ..., t_m \rangle$
- size |s|: number of itemsets in s
- length l(s): $l(s) = \sum_{i=1}^{n} |t_i|$

Preliminary Concepts

α sub-sequence of β OR β super-sequence of α (contains)

- $\alpha = \langle \alpha_1, \alpha_2, ..., \alpha_m \rangle$
- $\beta = \langle \beta_1, \beta_2, ..., \beta_m \rangle$
- $\alpha \sqsubseteq \beta$ (if $\alpha \neq \beta$, written as $\alpha \sqsubset \beta$)
- iff $\exists i_1, i_2, ..., i_m$, such that
  $1 \leq i_1 < i_2 < ... < i_m \leq n$ and
  $\alpha_1 \subseteq \beta_i, \alpha_2 \subseteq \beta_{i_2}, ..., \alpha_m \subseteq \beta_{i_m}$
- β absorbs α: if β contains α and their *support* are the same

Preliminary Concepts

Support

- $D = \{s_1, s_2, ..., s_n\}$: sequence database
- each s associated with id (id of s_i is i)
- $|D|$: number of s in D
- $support(\alpha)$: number of s in D which contain α
  $support(\alpha) = |\{s|s \in D \ and \ \alpha \sqsubseteq s\}|$
- min_sup: minimum support threshold

Preliminary Concepts

Frequent sequential pattern (FS) and closed FS (CS)

- FS: includes all s of $support(s) \leq min\_sup$
- $CS = \{\alpha | \alpha \in \; FS \; and \; \nexists \beta \in$ FS such that $\alpha \sqsubseteq \beta \; and \; support(\alpha) = support(\beta)\}$
- *closed sequence mining*: find CS above min_sup
- database containment relation $D \sqsubseteq D'$: if $\exists$ an injective function $f : D \to D'$, s.t. $\forall s \in D, s \sqsubseteq f(s)$