SuSe: Summary Selection for Regular Expression Subsequence Aggregation over Streams

Steven Purtzel
Humboldt-Universität zu Berlin, Germany
purtzesc@informatik.hu-berlin.de

Matthias Weidlich
Humboldt-Universität zu Berlin, Germany
matthias.weidlich@hu-berlin.de

ABSTRACT

Regular expressions (RegEx) are an essential tool for pattern matching over streaming data, e.g., in network and security applications. The evaluation of RegEx queries becomes challenging, though, once subsequences are incorporated, i.e., characters in a sequence may be skipped during matching. Since the number of subsequence matches may grow exponentially in the input length, existing RegEx engines fall short in finding *all* subsequence matches, especially for queries including Kleene closure.

In this paper, we argue that common applications for RegEx queries over streams do not require the enumeration of all *distinct* matches at *any* point in time. Rather, only an aggregate over the matches is typically fetched at specific, yet unknown time points. To cater for these scenarios, we present SuSe, a novel architecture for RegEx evaluation that is based on a query-specific summary of the stream. It employs a novel data structure, coined State-Summary, to capture aggregated information about subsequence matches. This structure is maintained by a summary selector, which aims at choosing the stream projections that minimize the loss in the aggregation result over time. Experiments on real-world and synthetic data demonstrate that SuSe is both effective and efficient, with the aggregates being based on several orders of magnitude more matches compared to baseline techniques.

PVLDB Reference Format:
Steven Purtzel and Matthias Weidlich. SuSe: Summary Selection for Regular Expression Subsequence Aggregation over Streams. PVLDB, 14(1): XXX-XXX, 2020.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/spurtzel/SuSe.

1 INTRODUCTION

The evaluation of regular expressions (RegEx) over streaming data is at the core of applications in various domains, such as network monitoring [18, 23, 26, 27], financial fraud detection [38, 42], or infrastructure security [8, 22]. RegEx have been studied extensively in theoretical computer science [17, 19] and automata-based approaches for their evaluation, most prominently the Thompson construction [40] and its derivates have been around for several decades.

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

Existing RegEx engines, such as PCRE2 [12], pcregrep [16], Oniguruma [1], Boost.Regex [24], RE2 [34] and TRE [41] are limited in their ability to consider subsequences during evaluation. That is, they lack native support for skipping characters in a sequence when constructing matches and do not support the enumeration of *all* matches. This is not a matter of expressiveness, though, since the skipping of characters can be encoded explicitly, e.g., by transforming a regular expression $\gamma = abc$ into $\gamma' = \Sigma^* a \Sigma^* b \Sigma^* c \Sigma^*$ with Σ being the underlying alphabet. Rather, the efficiency of the evaluation of regular expressions is the limiting factor, due to the exponential growth of the matches in the size of the input. REmatch [35] is a notable exception in that it supports the enumeration of *all* matches; yet, it suffers from performance issues due to enumerating exponentially many matches, as we demonstrate later empirically.

Complex Event Processing (CEP) engines like FlinkCEP [13] allow for subsequence matching using automaton-based methods. They evaluate sequence queries grounded in regular expressions, enhanced with predicates and time window constraints. However, these engines also face performance issues caused by the need to manage exponentially many matches. The CORE engine [6] seeks to address this by representing matches in an efficient data structure. Still, it realizes an exhaustive enumeration of matches, leading to significant performance bottlenecks, as demonstrated by our experiments.

In this paper, we argue that RegEx evaluation with subsequences may be optimized based on the following two observations:

(1) Some application scenarios demand that all matches of a regular expression are incorporated, whereas the output is given in terms of an aggregate (e.g., count or sum) computed over all matches. Such scenarios have been referred to as *event trend aggregation* and dedicated solutions that are based on incremental computation of the aggregate have been developed for it, e.g., GRETA [31] and its successors [28, 30, 32, 36].

(2) The matches of a regular expression, and aggregates over them, are not necessarily needed at any point in time. Rather, many application scenarios involve some external trigger that determines when the result is considered to be relevant. For instance, users may load an analysis report, open a dashboard, or refresh a visualization at certain, generally unknown time points.

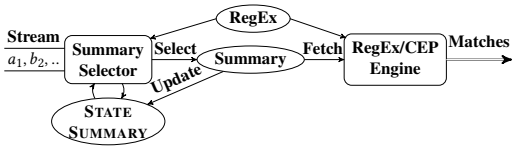


Figure 1: SuSe architecture for RegEx evaluation over streams.

Combining these two observations, we present SuSe, a novel architecture for RegEx evaluation that is based on a query-specific summary of the stream, which is illustrated in Figure 1. Specifically, SuSe makes the following contributions:

(1) We present a model for RegEx subsequence matching with aggregations that is based on a summary of a stream. Based thereon, we formulate the problem of constructing an optimal summary that minimizes the expected information loss.
(2) We introduce StateSummary, a novel data structure to capture aggregated information about subsequence matches. We elaborate on the operations to maintain this structure.
(3) We present an approach for summary selection, which aims at constructing an optimal summary. It is based on a cost model to assess the potential benefit of stream elements.

Below, we first provide a motivating example (§2) and introduce preliminaries (§3), before giving a formal problem statement (§4). Then, we define the StateSummary (§5), as the basis for summary selection (§6). We present evaluation results (§7), demonstrating that SuSe computes aggregates based on several orders of magnitude more matches than baseline techniques. Finally, we review related work (§8) and conclude the paper (§9).

2 MOTIVATING EXAMPLE

A bike rental service aims to enhance operational efficiency by analyzing patterns within trip data streams, e.g., targeting patterns of events denoting short trips in busy areas. These patterns, indicative of user behavior and operational dynamics, are defined using regular expressions. The inherent variability and complexity of real-world data streams, where a strictly contiguous occurrence of events within a pattern may be rare or too restrictive, motivates the search for subsequences, i.e., events in a pattern can occur in a non-contiguous order in the stream. Furthermore, streams are partitioned by attributes, e.g., bike ID or start location, to facilitate targeted analysis.

Rather than relying on continuous monitoring, which is resource-intensive, the service opts to generate summaries per partition that are queried at unknown points. Furthermore, to support detailed post-analysis, the service wants to retain a subsequence of the stream that led to the most pattern matches. This approach ensures the preservation of crucial information for in-depth analysis, while acknowledging a calculated trade-off: a minor information loss is accepted to maintain efficiency and storage feasibility.

Figure 2 showcases three strategies for retaining stream subsequences: Random, First-In-First-Out (FIFO), and SuSe. Given a stream s, a summary size n, a regular expression $r = AB^*C$, a time window $w = 10$, and an evaluation timestamp $e = 10$ (unknown beforehand), the goal is to select a subsequence $\mathcal{S}$ of s (i.e., an order-preserving projection) no larger than n, that maximizes the count of subsequence matches upon evaluating r against the summary $\mathcal{S}$.

Random. A first baseline strategy decides randomly, which of the elements is replaced by a new element, once the summary size n is reached. This way, a subsequence was chosen which produces zero matches upon evaluating r against $\mathcal{S}$, wasting the entire space.

FIFO. The FIFO strategy always keeps the latest n elements in the summary $\mathcal{S}$. Evaluating r against $\mathcal{S}$ generates two matches, i.e., A_7C_9 and $A_7B_8C_9$. However, keeping C_6 in the summary, which cannot contribute to any matches, illustrates inefficient space usage.

SuSe. Our SuSe comprises (1) the StateSummary, which efficiently captures how many matches $\mathcal{S}$ contains and how many matches each element in $\mathcal{S}$ participates in; and (2) an *efficient summary selection* function to decide which elements to keep. This way,

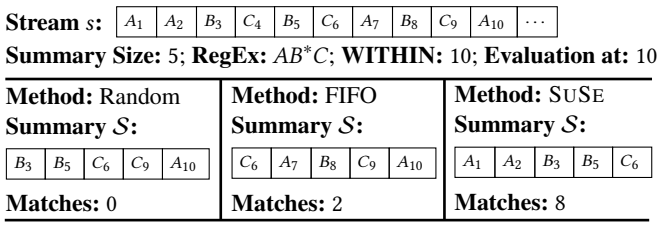


Figure 2: Comparison of random, First-In-First-Out (FIFO), and our SuSe summary selection strategy.

SuSe chooses a more effective stream summary, resulting in eight matches upon evaluating r over $\mathcal{S}$: A_1C_6; A_2C_6; $A_1B_3C_6$; $A_2B_3C_6$; $A_1B_5C_6$; $A_2B_5C_6$; $A_1B_3B_5C_6$; and $A_2B_3B_5C_6$. Note that, at time step 10, the total match count reaches 30 when evaluating r against s.

3 PRELIMINARIES

3.1 Regular Expressions

Syntax. An *alphabet* $\Sigma = \{a_1, \ldots, a_m\}$ is a totally ordered set of *characters* $a_1 < \ldots < a_m$, where $m \in \mathbb{N}^+$. A *valued character* c^v is an extension of a character, annotated with an attribute value $v \in \mathbb{N}$. The latter models a potential payload of a stream element and could also be defined over a different domain, depending on the application scenario. Most of our techniques are based on non-valued characters; however, for certain extensions, we will refer to valued characters. A *word* is a sequence of characters $w = \langle c_1, \ldots, c_n \rangle$ of size $n = |w|$ over Σ, i.e., $c_i \in \Sigma$. The concatenation of words $w = \langle c_1, \ldots, c_n \rangle$ and $w' = \langle c'_1, \ldots, c'_m \rangle$ is $w.w' = \langle c_1, \ldots, c_n, c'_1, \ldots, c'_m \rangle$. A *prefix* of w is any word of length $\leq n$ that starts from its first character.

A *subsequence* y of a word $w = \langle c_1, \ldots, c_n \rangle$, denoted by $y \preceq w$, is obtained by deleting zero or more characters from w without altering the order of the remaining characters. Hence, there are 2^n possible subsequences of w. Formally, a subsequence y is represented as $y = \langle c_{i_1}, \ldots, c_{i_k} \rangle$, where $k \leq n$ and $1 \leq i_1 \leq \ldots \leq i_k \leq n$.

To establish the relation between the positions in y and w, a *mapping function* $m \colon \{1, \ldots, k\} \to \{1, \ldots, n\}$ is defined such that $j \mapsto i_j$. This function maps the k positions of y to the corresponding n positions in w, denoted as $y \preceq_m w$. A *partial mapping* of a subsequence y' to w, denoted as $y' \preceq_{pm} w$, involves any prefix y' of y that can be mapped to w using function pm. Note that multiple (partial) mapping functions may exist from a subsequence to a word.

Example 1. *Consider the alphabet* $\Sigma = \{a, b, c, d\}$ *and a word* $w = \langle a_1, b_2, d_3, c_4, a_5, b_6, c_7, d_8 \rangle$. *Here,* $y = \langle a, b, d, c \rangle$ *is a subsequence with two different mappings,* $y \preceq_{m_1} w$ *with* $m_1 \colon 1 \mapsto 1, 2 \mapsto 2, 3 \mapsto 3, 4 \mapsto 4$*;* $y \preceq_{m_2} w$ *with* $m_2 \colon 1 \mapsto 1, 2 \mapsto 2, 3 \mapsto 3, 4 \mapsto 7$.

The character at the i-th position in word w is denoted by $w[i]$, where $1 \leq i \leq |w|$. The set of all words of length n over an alphabet Σ is Σ^n; and the set of all possible words is $\Sigma^* = \bigcup_{n \geq 0} \Sigma^n$. A *language* L over Σ is any subset $L \subseteq \Sigma^*$, and the set of all prefixes of L is denoted as L_{pre}. The *empty word* is $\varepsilon = \Sigma^0$.

Let A and B be languages defined over an alphabet Σ. Then, we define concatenation, union, and Kleene star for languages as:

- $AB = \{x.y \mid x \in A, y \in B\}$,
- $A \cup B = \{x \in \Sigma^* \mid x \in A \vee x \in B\}$, and
- $A^* = \bigcup_{n \geq 0} A^n$.

Semantics. We inductively define regular expression semantics. The symbols $\emptyset$, ε, and $a \in \Sigma$ are regular expressions, describing:

- the empty language $L(\emptyset) = \emptyset$;
- the language $L(\varepsilon) = \langle\varepsilon\rangle$, and
- for each character $a \in \Sigma$, the language $L(a) = \{\langle a\rangle\}$.

Let α and β be regular expressions, with languages $L(\alpha)$ and $L(\beta)$. Then, semantics of concatenation $\alpha\beta$, union $(\alpha|\beta)$, and Kleene star $(\alpha)^*$ are defined as follows:

- $L(\alpha\beta) = L(\alpha)L(\beta)$,
- $L(\alpha|\beta) = L(\alpha) \cup L(\beta)$, and
- $L((\alpha)^*) = L(\alpha)^*$.

3.2 Regular Expression Queries

Positive natural numbers represent time points, denoted as $(\mathbb{T}, \leq)$ where $\mathbb{T} \subseteq \mathbb{N}^+$. Each character c_i is associated with a timestamp $c_i.t \in \mathbb{T}$. Time progresses as ascending positive natural numbers. A *regular expression query* $q = (\gamma, \tau, \mathscr{A})$ is a triple, comprising a regular expression γ (also denoted as q_γ), a time window $\tau \in \mathbb{T}$ (also q_τ), and an aggregate function $\mathscr{A}$ (also $q_\mathscr{A}$). A mapping $y \preceq_m w$ of a subsequence $y \in L(\gamma)$ to a word $w \in \Sigma^*$ is a *match* of a regular expression query q, if the time window constraint is satisfied, i.e., $w[m(k)].t - w[m(1)].t \leq q_\tau$.

Example 2. *For $\Sigma = \{a, b, c\}$, consider a non-valued character stream $s = \langle a_1, b_2, c_3, c_4\rangle$ and a time window $q_\tau = 2$. For $q_\gamma = abc$, two possible mappings exist: $m_1 \colon 1 \mapsto 1, 2 \mapsto 2, 3 \mapsto 3$ and $m_2 \colon 1 \mapsto 1, 2 \mapsto 2, 3 \mapsto 4$. Due to q_τ, only m_1 is a match.*

Conversely, a *partial match* occurs when a subsequence $y \in L_{\text{pre}}(\gamma)$ maps to positions of a word. Such a partial match is a candidate for becoming a complete match as more characters of w are examined. Notably, a complete match can also serve as a partial match. For instance, for $\gamma = a(b|c)d^*$, any match of the word $\langle a, b, d\rangle \in L(\gamma)$ also serves as a partial match of $\langle a, b, d, d\rangle \in L(\gamma)$.

Our work utilizes a streaming model of computation. Initially, the regular expression query $q = (\gamma, \tau, \mathscr{A})$ is received and preprocessed. After preprocessing, the aggregate function $\mathscr{A}$ dictates whether we receive a stream of non-valued characters $s = \langle c_1, c_2, \dots\rangle$ or valued characters $s^v = \langle c_1^{v_1}, c_2^{v_2}, \dots\rangle$, where $c_i \in \Sigma$ and $v_i \in \mathbb{N}$.

We focus on the aggregate functions $\mathscr{A} \in \{\text{Count}, \text{Sum}, \text{Avg}\}$, which are defined over the matches of a regular expression query q. Count calculates the number of matches. Sum and Avg apply only to a stream of valued characters; Sum adds up the values attached to characters of all matches, and Avg is the ratio of Sum to Count.

3.3 Regular Expression Query Matching

Automata-Based Matching. Traditional matching of a regular expression q_γ against a stream (word) relies on a Non-deterministic Finite Automaton (NFA). An NFA is a tuple $\mathcal{N} = (Z, \Sigma, \delta, z_0, F)$, with Z as a finite set of states, Σ as the alphabet, $z_0 \in Z$ as the initial state, $\delta \colon Z \times \Sigma \to 2^Z$ as the transition function, and $F \subseteq Z$ as the set of final states. The NFA for q_γ is obtained by the *Thompson construction* [40] that recursively composes NFAs of sub-expressions, such that all words in $L(q_\gamma)$ are accepted.

As motivated, we aim to detect *all possible (partial) matches of subsequences from $L(q_\gamma)$ to positions in a word w*. This mirrors the most challenging combination of consumption and selection policies [2, 14, 43] in complex event processing: *Reuse* and *Skip-Till-Any-Match (STAM)*. A simple automata-based matching method works as follows: Each incoming character is checked against the current set of automata. If the character enables a transition for an automaton, it is duplicated, with one copy disregarding the transition and the other one taking it, yielding an exponential growth of the number of automata in the number of processed characters.

Table 1: Overview of notations for regular expressions.

Notation	Explanation
$\Sigma = \{a_1, \dots, a_m\}$	Alphabet with $a_1 < \dots < a_m$ characters.
c	A *non valued* character $c \in \Sigma$.
c^v	A *valued* character $c \in \Sigma$ with value $v \in \mathbb{N}$.
$w = \langle c_1, \dots, c_n\rangle$	Word (sequence of characters) of size $n = \lvert w\rvert$.
$w[i]$, $w[i].t$	Character at i-th position of w and its timestamp t.
$y \preceq_m x$	A subsequence y with a mapping m onto the string x.
$q = (\gamma, \tau, \mathscr{A})$	Regular expression query q, composed of regular expression γ, a time window size τ, and an aggregate function $\mathscr{A}$.
$s = \langle c_1, c_2, \dots\rangle$	A *stream* of *non valued* characters.
$s^v = \langle c_1^{v_1}, c_2^{v_2}, \dots\rangle$	A *stream* of *valued* characters.
$L(\gamma), L_{\text{pre}}(\gamma)$	The language of *all words*, the language of *all prefixes* of γ.
PM_w^q, CM_w^q	The sets of partial and complete matches respectively, resulting from evaluating q over $w \in \Sigma^*$.

Regular Expression Subsequence Matching. To formalize the evaluation of a regular expression query $q = (\gamma, \tau, \mathscr{A})$, only q_γ and q_τ are required, as the aggregate function $\mathscr{A}$ is applied to the resulting matches. We capture the matches resulting from evaluating q over a word $w \in \Sigma^*$ using two sets: partial matches PM_w^q and complete matches CM_w^q. These sets are based on the language $L(q_\gamma)$ and its prefix language $L_{\text{pre}}(q_\gamma)$. PM_w^q includes a partial match if it maps a subsequence $y \in L_{\text{pre}}(q_\gamma)$ to positions in w. CM_w^q includes a match if it maps a subsequence $y \in L(q_\gamma)$ to positions in w. Either way, the mapping must satisfy the time window q_τ.

Next, we formalize the construction of CM_w^q, while the construction of PM_w^q works analogously. Let $n = \lvert w\rvert$ and $[n] = \{1, \dots, n\}$. We define $\mathcal{M}_{\text{CM}} = \{\varphi \mid \varphi \colon L(q_\gamma) \to 2^{[n]}\}$ as the set of all *word mapping functions* φ from subsequences in $L(q_\gamma)$ to sets of up to n positions in w. Note that the mapping function m introduced earlier can be depicted through the word mapping function φ_m, e.g., in Example 2, m_1 aligns with $\varphi_{m_1}(\langle a, b, c\rangle) = \{1, 2, 3\}$. Analogously, the set of all *partial word mapping functions*, $\mathcal{M}_{\text{PM}} = \{\varphi' \mid \varphi' \colon L_{\text{pre}}(q_\gamma) \setminus L(q_\gamma) \to 2^{[n]}\}$, is based on $L_{\text{pre}}(q_\gamma)$.

A word mapping function φ for a subsequence $y \in L(q_\gamma)$, represented by $\varphi(y) = \{p_1, p_2, \dots, p_{|y|}\}$ with $p_1 < p_2 < \dots < p_{|y|}$, is deemed *valid* if it satisfies the following conditions:

(1) (Reconstruction) The positions $\varphi(y) = \{p_1, p_2, \dots, p_{|y|}\}$ in w must reconstruct $y \in L(q_\gamma)$, i.e., $y = w[p_1]w[p_2]\dots w[p_{|y|}]$.
(2) (Time Window Constraint) The mapping must adhere the time window constraint, i.e., $w[p_{|y|}].t - w[p_1].t \leq q_\tau$.

Based thereon, we define the set $CM_w^q = \{\varphi \mid \varphi \in \mathcal{M}_{\text{CM}}, \varphi \text{ is valid}\}$ as all valid word mapping functions φ. Analogously, PM_w^q denotes the set of all valid partial word mapping functions φ'.

Finally, using the set of complete matches as CM_w^q, we integrate the aggregate function $\mathscr{A}$. With $q_\mathscr{A} = \text{Count}$, the result is $\lvert CM_w^q\rvert$, i.e., the total number of complete matches. If $q_\mathscr{A} = \text{Sum}$, the result is $\sum_{\varphi \in CM_w^q} \sum_{i \in \text{Range}(\varphi)} w[i]^v$, summing all attribute values v of complete matches. For $q_\mathscr{A} = \text{Avg}$, the result is $\text{Sum}(CM_w^q)/\text{Count}(CM_w^q)$.

EXAMPLE 3. *Consider an alphabet* $\Sigma = \{a, b, c\}$ *and a regular expression query* $q = (\gamma, \tau, \mathscr{A})$ *with* $q_\gamma = abc$ *and* $q_\tau = 5$. *The language of* q_γ *is* $L(q_\gamma) = \{\langle a, b, c\rangle\}$. *For the stream* $s^v = \langle a_1^2, c_2^3, b_3^5, a_4^6, c_5^{13}, b_6^7, c_7^9\rangle$, *potential matches are* $\{\varphi_1 \colon \langle a, b, c\rangle \mapsto \{1, 3, 5\}, \varphi_2 \colon \langle a, b, c\rangle \mapsto \{1, 3, 7\}, \varphi_3 \colon \langle a, b, c\rangle \mapsto \{4, 6, 7\}\}$. *Incorporating the time window* q_τ, *we arrive at* $CM_{s^v}^q = \{\varphi_1 \colon \langle a, b, c\rangle \mapsto \{1, 3, 5\}, \varphi_3 \colon \langle a, b, c\rangle \mapsto \{4, 6, 7\}\}$. *For some aggregations* $q_{\mathscr{A}}$, *we get* $\textsc{Count}(CM_{s^v}^q) = 2$ *and* $\textsc{Sum}(CM_{s^v}^q) = (2+5+13)+(6+7+9) = 42$.

4 PROBLEM STATEMENT

Using the above model, we formulate the problem addressed in our work. We consider scenarios, in which the query evaluation is relevant at a random time instance (e.g., when a user refreshes a report). Let $U \colon \Omega \to \mathbb{T}$ be a random variable representing this trigger, with Ω as the sample space encompassing all potential trigger instances. The probability distribution of U is denoted by $P_U(t)$, specifying the probability that U equals t. Let $\mathcal{E} \subseteq \mathbb{T}$ be the set of *evaluation timestamps*, which are drawn independently as a fixed number N of samples from $\mathbb{T}$ according to P_U, yielding $\mathcal{E} = \{t_1, \ldots, t_N\}$. Each $t_i \in \mathcal{E}$ denotes a time instance when query results are relevant.

For a stream of valued characters $s^v = \langle c_1^{v_1}, c_2^{v_2}, \ldots\rangle$ over Σ, let $s_t^v = \langle c_1^{v_1}, \ldots, c_t^{v_t}\rangle$ denote the stream up to time instance $t \in \mathbb{T}$. We define a summary representation function $r_{\mathcal{S}}$ conceptually as $r_{\mathcal{S}}(s_t^v, n) = \langle c_{i_1}^{v_{i_1}}, \ldots, c_{i_k}^{v_{i_k}}\rangle$ with $1 \le i_1 \le \cdots \le i_k \le t$ and $k \le n$, where i_j are indices of the characters in s_t^v. For the stream s_t^v and the summary size n, this function provides a subsequence of the stream that consists of up to n characters of s_t^v. Technically, the summary representation function is realized by a summary selection algorithm that decides whether an element should be kept, discarded, or replaced, thus maintaining a subsequence of the stream.

We aim to assess how the projected characters in a summary affect a loss function $l_{\mathscr{A}} \colon \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ upon an evaluation trigger. The loss quantifies the discrepancy between the values x_{opt} and x, obtained by the aggregate for query matches over either the stream s_t^v (for x_{opt}) or the chosen subsequence $r_{\mathcal{S}}(s_t^v, n)$ (for x) at time t.

For aggregations COUNT and SUM, we employ the loss function $l_{\text{COUNT}}(x, x_{\text{opt}}) = l_{\text{SUM}}(x, x_{\text{opt}}) = |x - x_{\text{opt}}|$. For AVG, we define the loss as the relative error $l_{\text{AVG}}(x, x_{\text{opt}}) = |\frac{x - x_{\text{opt}}}{x_{\text{opt}}}|$. For each evaluation timestamp $t \in \mathcal{E}$, we aim to minimize the loss $l_{\mathscr{A}}$.

PROBLEM 1. *Let* Σ *be an alphabet and* $s_t^v = \langle c_1^{v_1}, \ldots, c_t^{v_t}\rangle$ *be a stream of valued characters up to time instance* $t \in \mathbb{T}$. *Also, let* $q = (\gamma, \tau, \mathscr{A})$ *be a regular expression query,* n *be the summary size,* $r_{\mathcal{S}}(s_t^v, n)$ *the summary representation function, and* $\mathcal{E} \subseteq \mathbb{T}$ *a finite set of evaluation timestamps drawn independently from a probability distribution* P_U. *Our objective is to find, for each time point* $t \in \mathcal{E}$, *the stream subsequence* $r_{\mathcal{S}}(s_t^v, n)$ *that minimizes the loss:*

$$\min_{r_{\mathcal{S}}(s_t^v, n)} (l_{\mathscr{A}}(q_{\mathscr{A}}(CM_{s_t^v}^q), q_{\mathscr{A}}(CM_{r_{\mathcal{S}}(s_t^v, n)}^q)))$$

$$\text{subject to: } |r_{\mathcal{S}}(s_t^v, n)| \le n, \ \forall\, t \in \mathcal{E}$$

5 STATE SUMMARY DATA STRUCTURE

The traditional continuous evaluation of regular expression queries over streams poses disadvantages for the problem introduced in §4. The number of automata that need to be maintained may grow exponentially in the number of processed characters, yielding an exponential runtime of an evaluation algorithm. Yet, many matches are materialized only to be discarded before the occurrence of an evaluation trigger, which wastes computational resources.

In the SUSE framework, we approach the evaluation of regular expression queries with a stream summary, as illustrated already in Figure 1. That is, the summary captures aggregated information about the stream, which, upon the occurrence of an evaluation trigger, enables the (approximate) computation of the aggregate over the complete matches of the query.

The core of SUSE is a stream summary and the STATESUMMARY data structure. Let $s^v = \langle c_1^{v_1}, c_2^{v_2}, \ldots\rangle$ be a valued (or non-valued) stream. Then, the summary $\mathcal{S} = \langle c_k^{v_k}, c_l^{v_l}, \ldots, c_m^{v_m}\rangle$ is a subsequence of the stream. Given a regular expression query q, the STATESUMMARY stores (i) aggregated information about partial and complete matches of q for the characters in $\mathcal{S}$ (i.e., *global state counters*); and (ii) information on the contribution of each character in $\mathcal{S}$ to these matches (i.e., *local state counters*). Moreover, this information is kept for an *active time window* that denotes a temporal context.

To simplify the presentation, we first introduce the STATESUMMARY for queries using COUNT as the aggregation function, in terms of its construction (§5.1) and maintenance (§5.2). Then, we discuss how it is adapted for SUM and AVG aggregations (§5.3).

5.1 STATESUMMARY Compilation

Given a regular expression query $q = (\gamma, \tau, \mathscr{A})$, the STATESUMMARY is constructed based on the NFA for q_γ obtained by the Thompson construction [40]. The NFA is further adapted by removing epsilon transitions, subsequent removal of unreachable states, and merging of states (which are final or non-final) with identical transitions. The resulting NFA, denoted as $\mathcal{N} = (Z, \Sigma, \delta, z_0, F)$, is used to compile our STATESUMMARY. The data structure includes state counters (summarized in Table 2) that are defined and derived as follows.

Global State Counter. For each state $z_i \in Z$ in $\mathcal{N}$, where $0 \le i < |Z|$, the *global state counter* $\#z_i \in \mathbb{N}$ keeps track of the number of (partial) matches in that state z_i, considering *all* selected characters $\mathcal{S}$. Initially, these counters are set to zero. Let $\mathcal{S}_{\#Z} = (\#z_0, \ldots, \#z_{|Z|})$ denote the tuple of *global state counters* of $\mathcal{S}$, derived from the states Z of the compiled NFA $\mathcal{N}$, where each component $\#z_i$ in $\mathcal{S}_{\#Z}$ denotes a global state counter. Referring to the example given in Figure 3, with global state counters $(\#z_0, \#z_1, \#z_2, \#z_3)$, processing the stream s yields 8 partial matches in z_1, 4 partial matches in z_2, and 10 complete matches in z_3, represented by the global state counters $(0, 8, 4, 10)$.

COUNT Rules. The COUNT rules specify the number of (partial) matches in a given state, based on the evaluation of query q over the selected characters $\mathcal{S}$. We derive them from the transitions of $\mathcal{N}$ by checking, for each state $z_i \in Z$ and each character $c \in \Sigma$, if there is a transition leading to any state in Z. As an illustration, consider the NFA in Figure 3, where $\delta(z_1, C) = \{z_1\}$, while $\delta(z_1, A) = \emptyset$.

With each character c, any (partial) match in a state z_i satisfying $\delta(z_i, c) \neq \emptyset$ can transition using c. This creates (partial) matches in the resulting state(s), as each partial match in z_i transitions using c, causing the initial count of partial matches in the subsequent state(s), given by $\delta(z_i, c)$, to increment by the current count z_i. Transitions from the initial state $z_0 \in Z$ to a state $z_k \neq z_0$ are special, as they represent the creation of a new (partial) match, adding one to z_k.

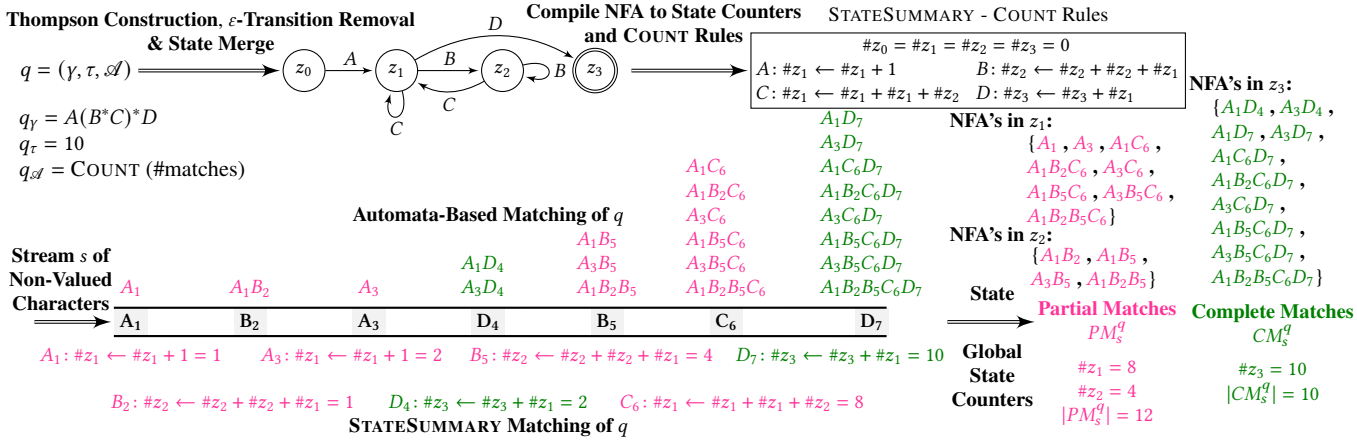


Figure 3: A non-valued character stream and the NFA for the regular expression $q_\gamma = A(B^*C)^*D$. The counters of the STATESUMMARY are contrasted with the partial matches constructed by traditional automata-based matching.

Table 2: Overview of different counter concepts.

Notation	Explanation
$\mathcal{S}_{\#Z}$	**Global state counters**: counting *all* (partial) matches in $\mathcal{S}$.
$c_{\#Z}$	**Local state counters**: counting (partial) matches c is involved.
$\mathcal{A}_{\#Z}$	**Active global state counters**: counting *all* (partial) matches satisfying q_τ, thus having the potential to lead to future matches.
$\mathcal{A}_{c_{\#Z}}$	**Active local state counters**: counting (partial) matches c is involved satisfying q_τ, thus having the potential to lead to future matches.

Based on $\mathcal{N}$'s transitions, we define COUNT rules to quantify the number of (partial) matches within a state resulting from matching an element against the current state. For each state $z_i \in Z$, global state counter $\#z_i$ in $\mathcal{S}_{\#Z}$, and character $c \in \Sigma$:

(1) If z_i has a self-loop with c:
if $\delta(z_i, c) = \{z_i\}$ then $\#z_i \leftarrow \#z_i + \#z_i$.

(2) If transitions from states $\{z_{i_1}, z_{i_2}, \ldots, z_{i_l}\}$ lead to z_k upon processing character c (excluding the self-loop case):
if $\delta(z_{i_j}, c) = \{z_k\}$ for $j \in \{1, ..., l\}$ then $\#z_k \leftarrow \#z_k + \sum_{j=1}^{l} \#z_{i_j}$.

(3) If the transition goes from the initial state z_0 to $z_k \neq z_0$:
if $\delta(z_0, c) = \{z_k\}$ then $\#z_k \leftarrow \#z_k + 1$.

Note that for counters impacted by both, a self-loop and a transition from a different state, the number must first be doubled by the self-loop. This way, we ensure that no (partial) match undergoes two transitions, maintaining the correctness of the counter.

Since δ yields a subset of 2^Z, a character c may update multiple state counters, with $O(|Z|)$ as an upper bound.

EXAMPLE 4. *Consider the stream $s = \langle A_1, B_2, A_3, D_4, B_5, C_6, D_7 \rangle$ in Figure 3. The expression $\gamma = A(B^*C)^*D$ yields global state counters $\mathcal{S}_{\#Z} = (\#z_0, \#z_1, \#z_2, \#z_3)$, each initialized to zero. The* COUNT *rules are depicted in the upper box in Figure 3. Processing s, A_1 increments $\#z_1$ by one. Then, B_2 increases $\#z_2$ by the sum of $\#z_1$ and $\#z_2$; A_3 adds one to $\#z_1$; and D_4 augments $\#z_3$ to 2. After processing s, it holds that $|PM_s^q| = \#z_1 + \#z_2 = 8 + 4 = 12$ and $|CM_s^q| = \#z_3 = 10$.*

Local State Counter. *Local state counters* enrich the information captured for characters in $\mathcal{S}$. That is, for each character c in $\mathcal{S}$, these counters, denoted as $c_{\#Z} = (c_{\#z_0}, \ldots, c_{\#z_{|Z|}})$ and initialized to zero, quantify the number of partial and complete matches involving character c across each state.

Local state counters are updated using the COUNT rules introduced above for global state counters. When a character c arrives, the local state counters $c_{\#Z}$ reflect the (partial) matches that c contributes to the global state counters; thus, $c_{\#Z}$ is updated first based on the values of the global state counters in $\mathcal{S}_{\#Z}$. Subsequent updates of $c_{\#Z}$, however, rely on the values in their respective local state counters, to incorporate new (partial) matches involving c.

EXAMPLE 5. *Consider Figure 3 and character B_5. Upon its arrival, B_5 is incorporated by all partial matches in states z_1 and z_2. Hence, we have $B_{5\#z_2} \leftarrow \#z_2 + \#z_1 = 3$. Note that only when B_5 is added, we apply the global state counters $\#z_i$ in $\mathcal{S}_{\#Z}$ to the local state counters. Afterwards, the* COUNT *rules update B_5's local state counters by factoring in the corresponding local state counter values. Therefore, when C_6 is processed next, the local state counter for B_5 gets updated: $B_{5\#z_1} \leftarrow B_{5\#z_1} + B_{5_{\#z_1}} + B_{5\#z_2} = 3$.*

If each state is reachable from any other state, for each character c, we must monitor all $|Z|$ counters, which yields a space complexity for both types of counters of $O(|Z| + |\mathcal{S}| \cdot |Z|)$, which simplifies to $O(|\mathcal{S}| \cdot |Z|)$, where typically $|Z| \ll |\mathcal{S}|$.

Sliding Time Window. The time window q_τ of a regular expression query constrains matches based on the timestamps of characters. If a character c is added to the summary, it is possible that not all its elements fall within the same time window as c. Hence, the global state counters $\mathcal{S}_{\#Z}$ may not represent the correct values for initializing the local state counters $c_{\#Z}$ for character c. We now detail the extensions to STATESUMMARY for such temporal context.

We define the *active time window* $\mathcal{A}$ as the subsequence of the selected characters $\mathcal{S}$ that, upon adding a character c, satisfy the time window q_τ with c. Based thereon, we maintain separate *active* global and local state counters, denoted by $\mathcal{A}_{\#Z}$ and $\mathcal{A}_{c_{\#Z}}$, respectively. They reflect the match counts per character in $\mathcal{A}$ and are updated based on the same COUNT rules as before. Within $\mathcal{A}$, the term $\mathcal{A}_{init}$ denotes the oldest initiator element, i.e., the oldest character selected from the stream that aligns with a transition in the automaton starting in the initial state z_0. Note that there is no need to keep characters before this element that do not align with transitions from the initial state since they cannot contribute to new (partial) matches.

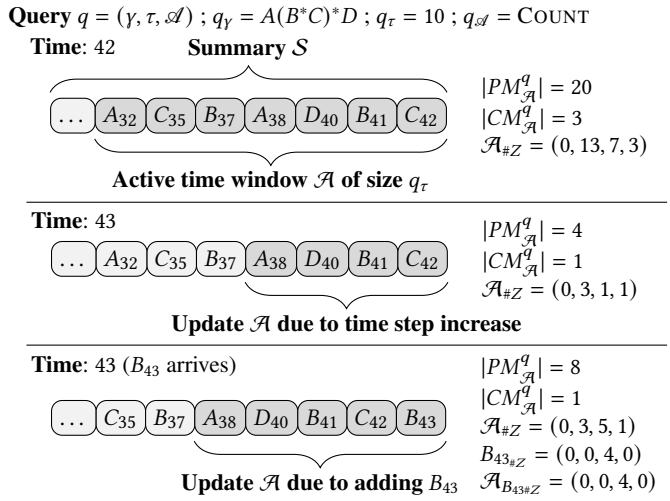


Figure 4: Evolution of a summary for the given regular expression query; shaded elements are in the active time window $\mathcal{A}$.

EXAMPLE 6. *Consider the summary $\mathcal{S}$ at time 42 in Figure 4 and the query q from Figure 3. Here, C_{42} is the most recently appended character in the summary. Owing to q_τ, A_{32} remains the earliest initiator that still matches C_{42}. Character $\mathcal{A}_{init} = A_{32}$ remains in $\mathcal{A}$, as it satisfies the time window. Yet, at time 43, it is no longer valid and hence is removed from $\mathcal{A}$, but it remains in $\mathcal{S}$. Both C_{35} and B_{37} fit within the time window. Since they are no initiators, they are discarded from $\mathcal{A}$ and we set $\mathcal{A}_{init} = A_{38}$.*

As we add B_{43}, its local state counters $B_{43\#Z}$ are initialized from the active time window state counters $\mathcal{A}_{\#Z} = (0, 3, 1, 1)$, so that we arrive at $B_{43\#Z} = (0, 0, \mathcal{A}_{\#z_1} + \mathcal{A}_{\#z_2}, 0) = (0, 0, 4, 0)$. These counts represent the current number of (partial) matches. Then, we update the global state counter values by adding the contributions from B_{43}: $\mathcal{S}_{\#Z} \leftarrow \mathcal{S}_{\#Z} + B_{43\#Z}$, meaning $\mathcal{S}_{\#z_i} \leftarrow \mathcal{S}_{\#z_i} + B_{43\#z_i}$ for $0 \le i < |Z|$.

In addition, we maintain *active local state counters* $\mathcal{A}_{c_{\#Z}}$ for characters c in $\mathcal{A}$ to track the number of matches each element participates in within $\mathcal{A}$ (whereas the local state counters $c_{\#Z}$ also cover characters no longer in $\mathcal{A}$). This nuanced tracking is crucial to later estimate the benefit of a character in $\mathcal{A}$ to yield future matches.

As soon as a character c drops out of $\mathcal{A}$, its local state counters $c_{\#Z}$ can no longer increase, but only decrease. The reason being that c can no longer lead to (partial) matches.

Maintaining the active global and local state counters results in additional space requirements of size $O(|\mathcal{A}| \cdot |Z|)$. Since $|\mathcal{A}|$ depends on q_τ and $|\mathcal{S}|$, we arrive at $O(min\{q_\tau, |\mathcal{S}|\} \cdot |Z|)$.

5.2 Operations

In this section, we detail the INITIALIZE, INSERT, and REMOVE operations that manipulate the summary $\mathcal{S}$.

INITIALIZE. The STATESUMMARY is initialized for a regular expression query $q = (\gamma, \tau, \mathscr{A})$, the NFA $\mathcal{N} = (Z, \Sigma, \delta, z_0, E)$ derived for it, and a user-defined summary *size* (n from Problem 1). We create tuples $\mathcal{S}_{\#Z}, \mathcal{A}_{\#Z}$ to capture the global state counters and the active global state counters, respectively. Note that the active time window $\mathcal{A}$ is a subsequence of $\mathcal{S}$, so that we do not store it explicitly.

Additionally, we allocate $O(size \cdot |Z|)$ memory for all future characters and their local state counters. Consequently, the runtime of the INITIALIZE operation is dominated by the initialization of the local state counter tuples for all future elements, which requires $O(|\mathcal{S}| \cdot |Z|)$ time and space.

We use an array to hold the characters when implementing the STATESUMMARY. While this choice induces an overhead for the REMOVE operation, it facilitates efficient linear scans, which are required by our INSERT and REMOVE operations.

INSERT. The INSERT(c) operation appends an element c from the stream to the summary $\mathcal{S}$. To incorporate c, it is necessary to adjust the global state counters $\mathcal{S}_{\#Z}$, the active global state counters $\mathcal{A}_{\#Z}$, and for each element e in $\mathcal{A}$ its active local state counters $\mathcal{A}_{e_{\#Z}}$ and local state counter $e_{\#Z}$. Finally, we have to initialize the active local state counter $\mathcal{A}_{c_{\#Z}}$ of c and its local state counters $c_{\#Z}$.

To comprehend the influence of character c on the underlying NFA $\mathcal{N}$, we assess its impact on ongoing (partial) matches within the active time window $\mathcal{A}$. Specifically, we identify the states affected by c. This is determined by the COMPUTESTATECOUNTERCHANGE function (line 11), which first initializes a new state counter tuple $newCounters$ of size $|Z|$ with zeros. Then, for each source state $from\_z \in Z$, we determine each target state $to\_z \in Z$, resulting from a transition using c, to increase the state counter $newCounters_{\#to\_z}$ by the count of $counters_{\#from\_z}$ (line 15). The resulting update tuple is saved to $globCounterChange$ (line 1).

In line 2 and line 3, we update the global state counters $\mathcal{S}_{\#Z}$ and active global state counters $\mathcal{A}_{\#Z}$ using $globCounterChange$ (i.e., by adding components of tuples, line 17). From line 4 to line 7, for each character e in the active time window $\mathcal{A}$, we determine the impact of c on the active local state counters of e, denoted by $localChange$ (line 5), to update its active local state counters $\mathcal{A}_{e_{\#Z}}$ and local state counters $e_{\#Z}$ by $localChange$. Finally, after appending the new character c to $\mathcal{S}$ (line 8), we initialize its active local state counters $\mathcal{A}_{c_{\#Z}}$ (line 9) and local state counters $c_{\#Z}$ (line 10).

The time complexity of INSERT is dominated by the updates of the (active) local state counters of characters in $\mathcal{A}$. The size of $\mathcal{A}$ is dictated by q_τ and, if $q_\tau > |\mathcal{S}|$, a single insert can affect *all* characters in $\mathcal{S}$. Since COMPUTESTATECOUNTERCHANGE requires $O(|Z|^2)$ time, it follows that the operation requires $O(min\{q_\tau, |\mathcal{S}|\} \cdot |Z_{\mathcal{N}}|^2)$ time using additional space $O(|Z|)$.

REMOVE. The REMOVE(c) operation (algorithm 2) removes element c from $\mathcal{S}$, which updates all counters. From global state counters $\mathcal{S}_{\#Z}$, we subtract the local state counters $c_{\#Z}$ (line 1), representing all (partial) matches involving c. Similarly, the active global state counters $\mathcal{A}_{\#Z}$ are reduced by the active local state counters $\mathcal{A}_{c_{\#Z}}$ (line 2), which signify the active (partial) matches involving c. When removing c, related (active) local state counters for remaining elements may also require updating.

The STATESUMMARY provides the count of (partial) matches for each element and state but lacks information regarding the joint matches between two characters c_1, c_2 from $\mathcal{S}$. The removal of c_1 may necessitate adjustments to the (active) local state counters of other elements in $\mathcal{S}$ if they had (partial) matches with c_1. Consequently, determining the proportion of joint (partial) matches in the respective counters $c_{1\#Z}$ and $c_{2\#Z}$ is required for updating $c_{2\#Z}$.

Algorithm 1: INSERT.

input : Summary $\mathcal{S}$; character c to insert; NFA $\mathcal{N} = (Z, \Sigma, \delta, z_0, E)$; active time window $\mathcal{A}$

1 globCounterChange ← `computeStateCounterChange` ($\mathcal{A}_{\#Z}, \mathcal{N}, c$);
2 $\mathcal{S}_{\#Z}$ ← `addStateCounters` ($\mathcal{S}_{\#Z}$, *globCounterChange*, $\mathcal{N}$);
3 $\mathcal{A}_{\#Z}$ ← `addStateCounters` ($\mathcal{A}_{\#Z}$, *globCounterChange*, $\mathcal{N}$);
4 **for** e *in* $\mathcal{A}$ **do**
5 localChange ← `computeStateCounterChange` ($\mathcal{A}_{e_{\#Z}}, \mathcal{N}, c$);
6 $\mathcal{A}_{e_{\#Z}}$ ← `addStateCounters` ($\mathcal{A}_{e_{\#Z}}$, *localChange*, $\mathcal{N}$);
7 $e_{\#Z}$ ← `addStateCounters` ($e_{\#Z}$, *localChange*, $\mathcal{N}$);
8 $\mathcal{S} \leftarrow \mathcal{S}.c$;
9 $\mathcal{A}_{c_{\#Z}}$ ← globCounterChange;
10 $c_{\#Z}$ ← globCounterChange;
11 **function** `computeStateCounterChange` (*counters*, $\mathcal{N}, c$)
12 newCounters ← $0^{|Z|}$;
13 **for** *from_z* ∈ Z **do**
14 **for** *to_z* ∈ δ(*from_z*, c) **do**
15 newCounters$_{\#to_z}$ ← newCounters$_{\#to_z}$ + counters$_{\#from_z}$;
16 **return** newCounters;
17 **function** `addStateCounters` (*stateCounters1, stateCounters2,* $\mathcal{N}$)
18 addCounters ← $0^{|Z|}$;
19 **for** $z \in Z$ **do** addCounters$_{\#z}$ ← stateCounters1$_{\#z}$ + stateCounters2$_{\#z}$;
20 **return** addCounters;

Algorithm 2: REMOVE.

input : Summary $\mathcal{S}$; character c to remove; query $q = (\gamma, \tau, \mathscr{A})$ NFA $\mathcal{N} = (Z, \Sigma, \delta, z_0, E)$; active time window $\mathcal{A}$

1 $\mathcal{S}_{\#Z} \leftarrow \mathcal{S}_{\#Z} - c_{\#Z}$;
2 $\mathcal{A}_{\#Z} \leftarrow \mathcal{A}_{\#Z} - \mathcal{A}_{c_{\#Z}}$;
3 $\mathcal{S}_{\text{included}}$.INITIALIZE($q, 2 \cdot q_\tau$);
4 $\mathcal{S}_{\text{excluded}}$.INITIALIZE($q, 2 \cdot q_\tau - 1$);
5 affectedElements ← $\{e$ in $\mathcal{S} \mid |e.t - c.t| \leq q_\tau\}$;
6 **for** e *in affectedElements* **do**
7 $\mathcal{S}_{\text{included}}$.INSERT($e$);
8 **if** $e \neq c$ **then**
9 $\mathcal{S}_{\text{excluded}}$.INSERT($e$);
10 **for** e *in affectedElements* **do**
11 $e_{\#Z} \leftarrow e_{\#Z} - (\mathcal{S}_{\text{included}e_{\#Z}} - \mathcal{S}_{\text{excluded}e_{\#Z}})$;
12 **if** $|c.t - currentTimestamp| \leq q_\tau$ **then**
13 `replayActiveTimeWindow()`;
14 **function** `replayActiveTimeWindow()`
15 $\mathcal{S}_{\text{temp}}$.INITIALIZE($q, q_\tau$);
16 **for** e *in* $\mathcal{A}$ **do**
17 $\mathcal{S}_{\text{temp}}$.INSERT($e$);
18 **for** e *in* $\mathcal{A}$ **do**
19 $\mathcal{A}_{e\#Z} = \mathcal{S}_{\text{temp}e\#Z}$;

The immediate question that follows is if we can discern the mutual (partial) matches based on the values in $c_{1\#Z}$ and $c_{2\#Z}$. Unfortunately, this is impossible. The reason is that c_1 and c_2 can match many elements that are exclusively matched with either c_1 or c_2. Even if two counters correspond to the same state but originate from distinct elements, their values can diverge heavily, making them unsuitable for determining the number of joint (partial) matches. Thus, we will now present a procedure that determines the joint counts.

Let c_i represent the element we intend to remove from $\mathcal{S}$. Due to the time window q_τ, element c_i can only match elements within the interval $[i - q_\tau, i + q_\tau]$. Consequently, only the (active) local state counters of elements within the interval of size $2 \cdot q_\tau$ need to be updated. Thus, we execute two replays over the range $[i - q_\tau, i + q_\tau]$, initiating from $i - q_\tau$ and finishing at $i + q_\tau$ (line 6). These replays instantiate additional local state counters for each element within

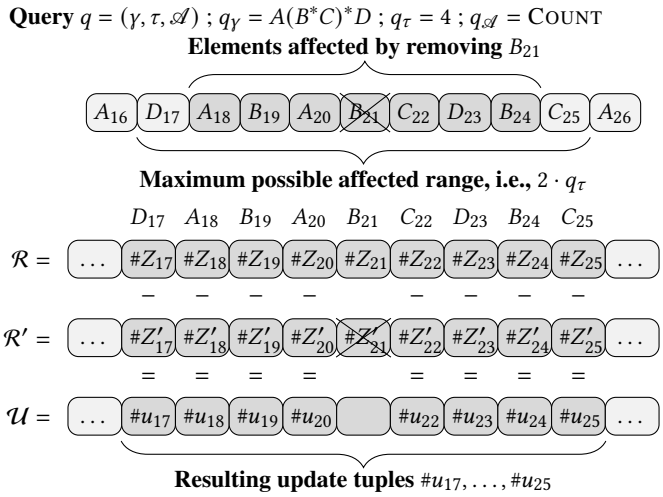


Figure 5: Evaluation of a summary when removing character B_{21}; for shaded elements, the local state counters are updated.

the interval, initializing them to zero. For updating the local state counters in the affected area, the initial replay incorporates c_i, while the latter omits c_i. Through this approach, we obtain two sets: $\mathcal{R}$, representing the local state counter tuples within the specified range post the first replay (c_i included) (line 7), and $\mathcal{R}'$, which stands for the local state counter tuples in the same range following post the second replay (c_i excluded) (line 9).

Next, we aim to determine the set of update tuples, $\mathcal{U}$, for the affected elements. For each update tuple $\#u \in \mathcal{U}$, the values within $\#u$ indicate the number of (partial) matches the corresponding element has in common with c_i for a given state z in Z. For an element c_j ($j \neq i$) in the affected range, we determine the corresponding replay tuples $\#v_j \in \mathcal{R}$ and $\#v_{j'} \in \mathcal{R}'$ such that the difference between $\#v_j - \#v_{j'}$ represents the shared match counts c_j and c_i have in common for each state. For instance, in Figure 5 for the element D_{17} the respective replay tuples are $\#Z_{17} \in \mathcal{R}$ and $\#Z_{17'} \in \mathcal{R}'$.

The update tuple is formulated as $\#u_j = \#v_j - \#v_{j'}$ and $\#u_j$ is then added to $\mathcal{U}$. Once $\mathcal{U}$ is computed, the local state counters, $c_{j\#Z}$, are adjusted for each affected c_j using the respective update tuple $\#u_j \in \mathcal{U}$ with the update operation $c_{j\#Z} \leftarrow c_{j\#Z} - \#u_j$ (line 11).

EXAMPLE 7. *Consider Figure 5, where $q_\tau = 4$ and we remove character B_{21}. Due to q_τ, the possible affected range $2 \cdot q_\tau$ needs revision, i.e., $[21 - q_\tau, 21 + q_\tau]$. We replay the affected range twice, first including B_{21}, obtaining the replay tuples $\mathcal{R}$, the second excluding B_{21}, obtaining $\mathcal{R}'$. Let $j \in \mathbb{N}$. When subtracting $\#R'_j$ in $\mathcal{R}'$ from $\#R_j$ in $\mathcal{R}$, we obtain the update tuples $\mathcal{U}$, with $\#u_j$ in $\mathcal{U}$ being the update tuple for $c_j \in \mathcal{S}$, which contains the joint counts of c_j and the removed character for each state counter. Based on these tuples, we update the local state counters of affected characters. If the removed character was present in $\mathcal{A}$, an additional replay of $\mathcal{A}$ yields updates to the active local state counters.*

Additionally, it is necessary to verify if the removed element falls within the active time window $\mathcal{A}$ (line 12). If so, *all* elements in $\mathcal{A}$ are potentially affected, making it necessary to update their active local state counters by replaying $\mathcal{A}$ separately once (line 13).

Table 3: Run time and space complexity of summary operations.

Operation	Time complexity	Space complexity
INITIALIZE	$O(\lvert\mathcal{S}\rvert \cdot \lvert Z\rvert)$	$O(\lvert\mathcal{S}\rvert \cdot \lvert Z\rvert)$
INSERT	$O(min\{q_\tau, \lvert\mathcal{S}\rvert\} \cdot \lvert Z\rvert^2)$	$O(\lvert Z\rvert)$
REMOVE	$O(min\{q_\tau, \lvert\mathcal{S}\rvert\}^2 \cdot \lvert Z\rvert^2)$	$O(min\{q_\tau, \lvert\mathcal{S}\rvert\} \cdot \lvert Z\rvert)$

We want to highlight two points: (1) To update the local state counters, more than a single replay of the affected area is required. Elements within the area can match those outside, creating shared counts that must be kept for correctness. To achieve this, we subtract the update tuples from the (active) local state counters. (2) The replay produces a zero tuple for elements in the area without matches to the removed element, as both replays yield identical tuples. For instance, this occurs for D_{17} in Figure 5 where $\#Z_{17} = \#Z'_{17}$.

As for the insertion, removing an element c_i could affect the local state counters of *all* elements in $\mathcal{S}$, if $q_\tau > |\mathcal{S}|$.

Complexity. The REMOVE operation implies a series of INSERT operations. For every new element inserted, all prior elements are updated. This results in a quadratic time complexity of $O(min\{q_\tau, |\mathcal{S}|\}^2 \cdot |Z|^2)$ and a space complexity of $O(min\{q_\tau, |\mathcal{S}|\} \cdot |Z|)$. We list an overview of our operations' time and space complexities in Table 3.

5.3 Model Extensions

Aggregate Functions. Having described the STATESUMMARY for queries involving COUNT as the aggregation, we turn to the changes needed to support SUM, which then also facilitates AVG.

As before, when a (valued) character c is added to $\mathcal{S}$, transitions of the automata are triggered, creating new (partial) matches in subsequent states. For all affected (partial) matches, the valued character c^v and its payload value v need to be incorporated. We define the SUM counters and rules based on those introduced for COUNT.

For each state $z_i \in Z$, we create a SUM counter $\mathcal{S}_{\#z_i} \in \mathbb{N}$, representing the current total sum over all payload values of (partial) matches in z_i. This yields a tuple of *global sum counters* $\mathcal{S}_{\#Z} = (\mathcal{S}_{\#z_0}, \ldots, \mathcal{S}_{\#z_{|Z|}})$. As for the COUNT rules, there are three cases we have to handle for defining the SUM rules: (1) self-loops, (2) transitions from other states, and (3) transitions originating from z_0: For each $z_i \in Z$, each $\mathcal{S}_{\#z_i}$ in $\mathcal{S}_{\#Z}$, and a valued character c^v:

(1) If z_i has a self-loop with c^v:
if $\delta(z_i, c^v) = \{z_i\}$ then $\mathcal{S}_{\#z_i} \leftarrow \mathcal{S}_{\#z_i} + \#z_i \cdot v$.

(2) If transitions from states $\{z_{i_1}, z_{i_2}, \ldots, z_{i_l}\}$ lead to z_k upon processing character c^v (excluding the self-loop case):
if $\delta(z_{i_j}, c^v) = \{z_k\}$ for $j \in \{1, 2, ..., l\}$ then
$\mathcal{S}_{\#z_k} \leftarrow \mathcal{S}_{\#z_k} + \sum_{j=1}^{l} \mathcal{S}_{\#z_{i_j}} + \#z_{i_j} \cdot v$.

(3) If the transition goes from the initial state z_0 to $z_k \neq z_0$:
if $\delta(z_0, c^v) = \{z_k\}$ then $\mathcal{S}_{\#z_k} \leftarrow \mathcal{S}_{\#z_k} + 1 \cdot v$.

The (active) local sum counters are defined analogously to their COUNT-based counterparts, but rely on the SUM rules.

Next, we will illustrate the SUM counter update process based on the SUM rules.

EXAMPLE 8 (STATESUMMARY MATCHING FOR SUM). *Consider Figure 3, where we will now set* $q_{\mathscr{A}}$ = SUM. *Due to the underlying NFA, we have the* SUM *counters* $\mathcal{S}_{\#Z} = (\mathcal{S}_{\#z_0}, \mathcal{S}_{\#z_1}, \mathcal{S}_{\#z_2}, \mathcal{S}_{\#z_3})$. *The resulting* SUM *rules are:*
$A^v : \mathcal{S}_{\#z_1} \leftarrow \mathcal{S}_{\#z_1} + 1 \cdot v$,
$B^v : \mathcal{S}_{\#z_2} \leftarrow \mathcal{S}_{\#z_2} + \mathcal{S}_{\#z_2} + \#z_2 \cdot v + \mathcal{S}_{\#z_1} + \#z_1 \cdot v$,
$C^v : \mathcal{S}_{\#z_1} \leftarrow \mathcal{S}_{\#z_1} + \mathcal{S}_{\#z_1} + \#z_1 \cdot v + \mathcal{S}_{\#z_2} + \#z_2 \cdot v$, *and*
$D^v : \mathcal{S}_{\#z_3} \leftarrow \mathcal{S}_{\#z_3} + \mathcal{S}_{\#z_1} + \#z_1 \cdot v$.
We modify the stream s *to* $s^v = \langle A_1^5, B_2^3, A_3^2, D_4^1, B_5^7, C_5^3, D_7^8 \rangle$. *After inserting* A_1^5, *we update* $\mathcal{S}_{\#z_1} \leftarrow 0 + 1 \cdot 5 = 5$. *When* B_2^3 *occurs,* $\mathcal{S}_{\#z_2} \leftarrow 0 + 0 \cdot 3 + 5 + 1 \cdot 3 = 8$ *and* A_3^2 *causes* $\mathcal{S}_{\#z_1} \leftarrow 5 + 1 \cdot 2 = 7$. *Due to* B_4^5, *we update* $\mathcal{S}_{\#z_2} \leftarrow 8 + 1 \cdot 4 + 7 + 2 \cdot 4 = 27$. *Encountering* C_5^7, *we update* $\mathcal{S}_{\#z_1} \leftarrow 7 + 2 \cdot 5 + 27 + 4 \cdot 5 = 64$. *Finally, we receive* D_7^8, *therefore,* $\mathcal{S}(\#z_3) \leftarrow 0 + 64 + 8 \cdot 8 = 128$, *resulting in* $\mathcal{S}(\#z_0, \#z_1, \#z_2, \#z_3) = (0, 64, 27, 128)$

Maintaining the STATESUMMARY for SUM and COUNT, support for AVG is obtained by using the ratio $\mathcal{S}_{\#z_i}/\#z_i$ for each state $z_i \in Z$.

Predicates. Patterns often materialize in stream partitions (e.g., per sensor, per person, per vehicle), making our model directly applicable for equality predicates over corresponding attribute values per partition. When defining predicates over subsets of query characters, e.g., $q_\gamma = ABC$ where B's and C's (but not A's) shall have equal attribute values, characters for which the predicate does not apply (here, the A characters) are replicated in each partition. Without partitioning, equality predicates can still be processed by maintaining counters for each attribute value previously considered for partitioning, tracking matches specifically for that attribute value.

Selection Policies. The semantics underlying our model is known as the STAM policy [2, 14, 43]. To integrate a stricter policy, such as Skip-Till-Next-Match (STNM) [14], we adapt the COUNT rules to reflect STNM's key distinction: it does not allow skipping *relevant* characters, i.e., (partial) matches are not duplicated. When a character arrives, it increases the count of (partial) matches in subsequent states by the number from the originating states. Under STNM, however, this increase is balanced by a decrease in the count of the originating states, as it does not permit match duplication. Applying STNM to our running example in Figure 3, when A_1 occurs, it increases $\#z_1$ by one. B_2 causes all partial matches in z_1 to transition to z_2, increasing $\#z_2$ by one. However, at the same time, $\#z_1$ is reduced by one (from one to zero) since an instance is not duplicated upon transition, which would reflect skipping the relevant character B_2.

6 SUMMARY SELECTION

To address Problem 1, we shall maintain a summary using the STATESUMMARY that aims at minimizing the loss in query evaluation. To this end, we first introduce a benefit function to quantify the potential of a character to contribute to a substantial amount of complete matches (§6.1), which then guides our selection strategy (§6.2). Again, we introduce the concepts for COUNT aggregations, and later discuss the changes required to support SUM and AVG (§6.3).

6.1 Benefit Function

To build a summary, i.e., a stream subsequence as introduced in §4, we define a benefit function $\mathcal{B}$ that scores a character based on the *present benefit* for current aggregation results, and the *expected benefit* for future results.

Present Benefit. The *present benefit* $\mathcal{B}_{\text{pres}}: \mathcal{S} \rightarrow \mathbb{N}$ (for COUNT) of character c in $\mathcal{S}$ is defined by the number of complete matches

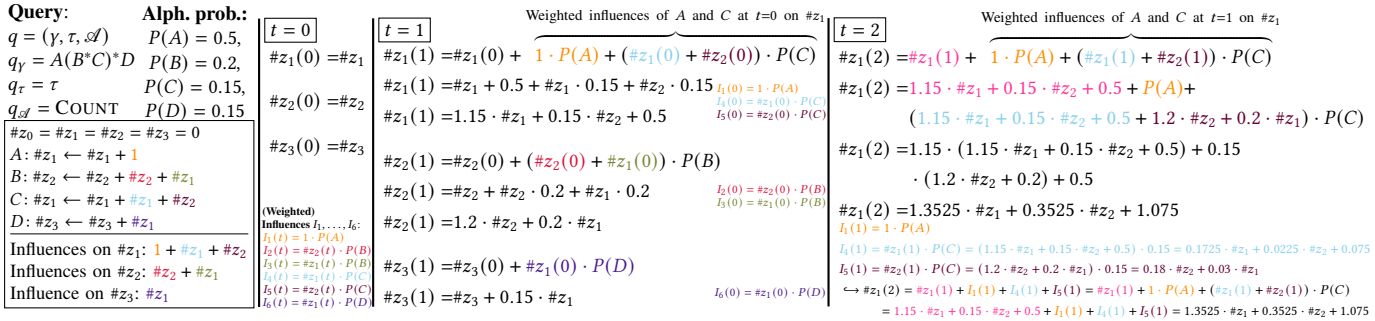


Figure 6: Extending the example from Figure 3, we compute the expected benefit for $t \in [0, 1, 2]$. Resulting coefficients are cached.

involving c, i.e., sum of local state counters for final states F in $\mathcal{N}$:

$$\mathcal{B}_{\text{pres}}(c) = \sum_{z_i \in F} c_{\#z_i} \tag{1}$$

We write $\mathcal{B}_{\mathcal{A}_{\text{pres}}}(c)$ to denote all complete matches of c within $\mathcal{A}$.

Expected Benefit. The *expected benefit* quantifies the expected count of complete matches with c using its active local state counter $\mathcal{A}_{c_{\#Z}}$ (i.e., (partial) matches with c that may lead to further matches) and the remaining time span Δ_τ, in which c can participate in matches.

We assume that the probability distribution over Σ, coined *alphabet probability* and denoted as $P_\Sigma \colon \Sigma \to [0, 1]$, is known. That is, $P_\Sigma(c)$ represents the probability that character c is the subsequent stream element. The time window q_τ determines the maximum time an element can lead to (partial) matches. As such, to compute the expected benefit, we induce for each time step $t \in [0, q_\tau]$ all possible characters $c \in \Sigma$ and measure their influence on the active state counters weighted by their occurrence probability $P_\Sigma(c)$. Based on the active local state counters $\#z_i$ in $\mathcal{A}_{c_{\#Z}}$, we compute how each $\#z_i$ is affected for each time step. For each $\#z_i$, the resulting count $\#z_i(t)$ represents the averaged count over all possible words Σ^t weighted by $P_\Sigma(c)$, i.e., the expected count of matches after t time steps.

Consider Figure 6, with the regular expression $q_\gamma = A(B^*C)^*D$ and the corresponding COUNT rules and their influences on state counters. Each rule models how a state counter is *influenced*. Since we create a COUNT rule for each transition in $\mathcal{N}$, the number of influences equals the number of transitions $T = \{(z, x, z') \mid z \in Z, x \in \Sigma, z' \in \delta(z, x)\}$. For each active state counter $\#z_i$ in $\mathcal{A}_{c_{\#Z}}$, we determine the *total influence* on $\#z_i$ for a single time step. That is, we determine all influences on $\#z_i$ corresponding to all transitions leading to z_i, i.e., $T_{z_i} = \{(z, x) \mid z \in Z, x \in \Sigma, z_i \in \delta(z, x)\}$. Each tuple $(z, x) \in T_{z_i}$ represents an influence on z_i. For instance, $T_{z_1} = \{(z_0, A), (z_1, C), (z_2, C)\}$, yielding the influences 1, $\#z_1$, and $\#z_2$, respectively, on $\#z_1$, which results in the total influence $1 + \#z_1 + \#z_2$. Moreover, each influence is weighted by probability $P_\Sigma(x)$ of character x occurring, e.g., for $\#z_1$, we get $1 \cdot P(A) + (\#z_1 + \#z_2) \cdot P(C)$.

For each transition in T, we define an *influence function*, $I_k \colon \mathbb{T} \to \mathbb{R}$, with $1 \leq k \leq |T|$, indicating the weighted influence of a tuple $(z, x, z') \in T$ at time t, yielding the set of all influences $\mathcal{I} = \{I_1, \ldots, I_{|T|}\}$. For every active state counter $\#z_i$ in $\mathcal{A}_{c_{\#Z}}$, we define an *influence set* $S_i \subseteq \{1, \ldots, |T|\}$, which contains the indices of influence functions that affect $\#z_i$. In Figure 6, $\#z_1$'s influence set is determined by $S_1 = \{1, 4, 5\}$, indicating that I_1, I_4, and I_5 affect $\#z_1$.

The computation at time $t + 1$ relies on the results at time t. As such, we derive a linear recurrence relation that operates on the active local state counters $\#z_i$ in $\mathcal{A}_{c_{\#Z}}$ of a character c in $\mathcal{S}$ to compute $c's$ the expected benefit and, therefore, the overall *benefit*.

Specifically, at time $t = 0$, we adopt $\#z_i(0) = \#z_i$ as the initial relation, signifying $\#z_i$'s current value. Then, the general linear recurrence relation for each $t + 1$ and counter $\#z_i$ is defined as:

$$\#z_i(t + 1) = \#z_i(t) + \sum_{k \in S_i} I_k(t) \tag{2}$$

To update $\#z_i(t + 1)$, we add to its previous value $\#z_i(t)$ the sum of weighted influences on $\#z_i$ at time t, e.g., at $t = 2$ in Figure 6, $\#z_1(2)$ is updated to $1.3525 \cdot \#z_1 + 0.3525 \cdot \#z_2 + 1.075$, calculated as $\#z_1(2) = \#z_1(1) + I_1(1) + I_4(1) + I_5(1) = \#z_1(1) + 1 \cdot P(A) + (\#z_1(1) + \#z_2(1)) \cdot P(C)$.

The output of Equation 2 relies on the initial active local state counters values $\#z_i$ in $\mathcal{A}_{c_{\#Z}}$ of c, which only increase. Hence, the result also captures c's active present benefit. Thus, to define the *expected benefit function* $\mathcal{B}_{\text{exp}} \colon \mathcal{S} \times \mathbb{T} \to \mathbb{R}$, we subtract out $\mathcal{B}_{\mathcal{A}_{\text{pres}}}(c)$:

$$\mathcal{B}_{\text{exp}}(c, t) = \sum_{z_i \in F} \mathcal{A}_{c_{\#z_i}}(t) - \mathcal{B}_{\mathcal{A}_{\text{pres}}}(c) \tag{3}$$

estimating the expected count of complete matches c leads to. For instance, let $\mathcal{A}_{c_{\#Z}} = (0, 5, 10, 15)$ in Figure 6; at $t = 0$, $\mathcal{A}_{c_{\#Z}}$ remains unchanged, while at $t = 1$, $\mathcal{A}_{c_{\#Z}}$ updates to $(0, 7.75, 13, 15.75)$.

Caching Coefficients. The computational cost for $\mathcal{B}_{\text{exp}}(c, t)$ grows as the time window q_τ increases. We therefore employ a cache that is built during pre-processing using the inputs P_Σ and $\mathcal{S}$. Using Figure 6 for illustration, we note that for each time step t and each active state counter $\#z_i$, a linear combination with constant coefficients is derived. A growing time window q_τ modifies only these coefficients, leaving the active state counter $\#z_i$ value anchored to its initial. Hence, for every $t \in [0, q_\tau]$, we store the coefficients corresponding to each state counter $\#z_i$. In Figure 6, for instance, for $\#z_1(2)$, the cached coefficients are 1.3525, 0.3525, and 1.075.

In total, for every $t \in [0, q_\tau]$, we keep track of all $O(|Z|)$ state counters by caching at most $O(|Z|)$ coefficients, leading to the memory requirement $O(q_\tau \cdot |Z| \cdot |Z|)$, simplifying to $O(q_\tau \cdot |Z|^2)$. After the cache is computed, for a given element c and t, we can compute $\mathcal{B}_{\text{exp}}(c, t)$ in time $O(|Z|)$ since we have to factor in all cached coefficients for a given input tuple.

6.2 Selection Strategy

The *selection strategy*, modelled as $\psi \colon \mathcal{S} \times \Sigma \to \mathcal{S}$, derives a new summary from the current summary upon the arrival of a character $c \in \Sigma$. It may insert c, replace c_{old} in $\mathcal{S}$ by c, or discard c (here,

$\mathcal{S}|c_{\text{old}}$ denotes the removal of character c_{old} from summary $\mathcal{S}$)

$$\psi(\mathcal{S}, c) = \begin{cases} \mathcal{S}.c & \text{if } c \text{ is inserted into } \mathcal{S} \\ (\mathcal{S}|c_{\text{old}}).c & \text{if } c_{\text{old}} \text{ is replaced by } c \\ \mathcal{S} & \text{if } c \text{ is discarded} \end{cases} \tag{4}$$

The decision of ψ relies on the benefit function $\mathcal{B}$. If the summary capacity has not been reached, a new character c is inserted. If the summary is filled already, the choice between replacement and discarding is done by iterating through the summary from the oldest to the newest element, identifying the character c' with the lowest benefit. For elements outside the active time window $\mathcal{A}$, it holds that $\mathcal{B}(c', 0) = \mathcal{B}_{\text{pres}}(c')$ is their benefit. Upon reaching $\mathcal{A}_{\text{init}}$, we set $i_{\text{old}} = i_{\text{new}} = \mathcal{A}_{\text{init}}$, marking both the oldest and initial newest initiators of $\mathcal{A}$. Note that i_{old} remains constant. As we proceed, encountering a new initiator updates i_{new}. We then determine values $\Delta_{\tau_{\text{old}}} = q_\tau - (c.t - i_{\text{old}}.t)$ and $\Delta_{\tau_{\text{new}}} = q_\tau - (c.t - i_{\text{new}}.t)$. The value $\Delta_{\tau_{\text{old}}}$ specifies the duration during which both c and i_{old} will be involved in matches. Conversely, $\Delta_{\tau_{\text{new}}}$ designates the overall period in which c will participate in matches at most. Based thereon, we define the *benefit function* $\mathcal{B}\colon \mathcal{S} \times \mathbb{T} \times \mathbb{T} \to \mathbb{R}$:

$$\mathcal{B}(c, \Delta_{\tau_{\text{old}}}, \Delta_{\tau_{\text{new}}}) = \mathcal{B}_{\text{pres}}(c) + \frac{\mathcal{B}_{\text{exp}}(c, \Delta_{\tau_{\text{old}}}) + \mathcal{B}_{\text{exp}}(c, \Delta_{\tau_{\text{new}}})}{2} \tag{5}$$

The function offers a balanced measure of the benefit of c, averaging its potential benefits over the periods $\Delta_{\tau_{\text{old}}}$ and $\Delta_{\tau_{\text{new}}}$.

The idea is based on the following observation: the value of $\Delta_{\tau_{\text{old}}}$ can be small, indicating that c might not contribute to matches for long; however, $\Delta_{\tau_{\text{new}}}$ dictates how long c can lead to matches. On the other hand, relying solely on $\Delta_{\tau_{\text{new}}}$ might not capture the expected benefit accurately, given that c aligns with i_{old} in many (partial) matches, but not for the span of $\Delta_{\tau_{\text{new}}}$.

Finally, we compute $c_{\#Z}$ and $\mathcal{A}_{c_{\#Z}}$ for c, to estimate its benefit $\mathcal{B}(c, \Delta_{\tau_{\text{old}}}, \Delta_{\tau_{\text{new}}})$. If the benefit of c' is smaller, we replace c' with c; otherwise, c is discarded. Since we compute $\mathcal{B}_{\text{exp}}$ in time $O(|Z|)$ for each element in $\mathcal{S}$, the run time complexity of the selection strategy ψ is given by $O(|\mathcal{S}| \cdot |Z|)$ using $O(q_\tau \cdot |Z|^2)$ space.

6.3 Modifications for Other Aggregate Functions

When adopting the SUM aggregate function, the benefit function needs to be adapted. The function for the present benefit is defined analogously, i.e., summing up all values of complete matches. Turning to the expected benefit function, we note that for COUNT, we estimate the count of matches that character c may generate. For SUM, each of these future matches carries a value v. Therefore, the expected benefit for c is augmented by its current present benefit (for SUM), combined with the expected match count of c weighted by v. While this function underestimates the expected SUM value for c, the underestimation is consistent over all characters in the summary.

7 EXPERIMENTAL EVALUATION

To evaluate SUSE, we conducted experiments using the setup described in §7.1. Specifically, we present results on the overall effectiveness, including an ablation and recall study (§7.2); on the sensitivity of the approach (§7.3); on a comparison against state-of-the-art engines for regular expressions and complex event processing (§7.4); and on two studies with real-world datasets (§7.5 and §7.6).

7.1 Experimental Setup

Our implementation and experimental setup is publicly available [39].

Datasets and Parameters. We used two real-world datasets and synthetic data to assess SUSE regarding the COUNT aggregate. First, we used one month of the Citi Bike dataset [9] that captures information about bike rentals. It contains ≈ 3.8 million events, each of them describing the duration of a trip, the start and end station, a bike ID, and information about the driver.

Our second real-world dataset, NASDAQ [25], contains stock trading data for 462028 events on a minute-by-minute basis. Each entry lists the stock symbol, opening, closing, highest and lowest prices, and the trading volume for that minute.

In synthetic data experiments, we assessed the influence of different parameters on SUSE, i.e., the summary size $|\mathcal{S}|$, stream size, number of evaluation timestamps $|\mathcal{E}|$, and the time window q_τ and length of q_γ of a query. To generate the data streams, characters were drawn from alphabets based on different distributions: Zipfian, normal, and uniform. The evaluation timestamps, see Problem 1, have been sourced from a Poisson process and a uniform distribution. Yet, the differences between both models have been negligible.

Baselines. We evaluated the performance of our approach against two baselines: A random baseline replaces a randomly chosen character in the summary once its capacity is met. The FIFO baseline consistently replaces the oldest element once its capacity is met.

Our ablation study used a restricted SUSE model, where the benefit function $\mathcal{B}$ is solely based on the present benefit $\mathcal{B}_{\text{pres}}$.

We examined STATESUMMARY's efficiency by comparing it to the state-of-the-art RegEx engine REmatch [35] and CEP engines FlinkCEP [13] and CORE [6]. For REmatch and CORE, we conducted experiments without outputting matches to minimize overhead and ensure fairness. In our Flink setup, while match printing was required to measure processing time for implementation reasons, subsequent evaluations confirmed a negligible impact on the results.

Metrics. We assess the evaluation quality using the *relative recall improvement*. At each evaluation time point $t \in \mathcal{E}$, we divide the current number of complete matches from SUSE by the current number of complete matches from a corresponding baseline. The average of the sum of these ratios denotes the relative recall improvement.

We also evaluate the *absolute recall*. At each evaluation timestamp, it is calculated by dividing the match count from SUSE by the ground truth, obtained by setting the summary size $|\mathcal{S}|$ equal to the stream size (limited to 2000 in these experiments). We also examine the *detected matches recall*, i.e., the ratio of complete matches that were present in SUSE to *all* potential matches over time.

We further report standard performance metrics, such as *throughput* (average number of elements processed per second), *latency* (average processing duration per element), and *memory usage* (maximum resident set size).

Implementation. Our experiments were conducted using a *C++* implementation, with 128-bit counters. They ran on a server with 4 Intel Xeon E7-4880 (60 cores, 1TB RAM).

7.2 Overall Effectiveness

We first compare against a random baseline and a FIFO baseline. In Figure 7a- 7b, we assessed how variations in the summary size $|\mathcal{S}|$

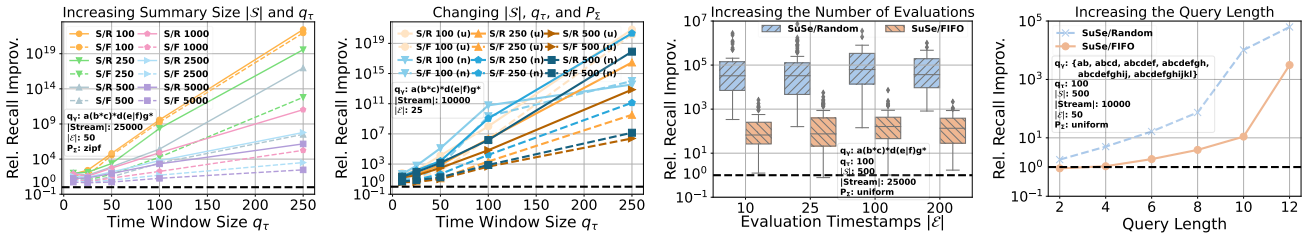


Figure 7: Examining the impact on relative recall improvement (higher is better) by varying (a) summary and time window sizes, (b) the alphabet probability distribution P_Σ, (c) the number of evaluation timestamps $|\mathcal{E}|$, and (d) the query length.

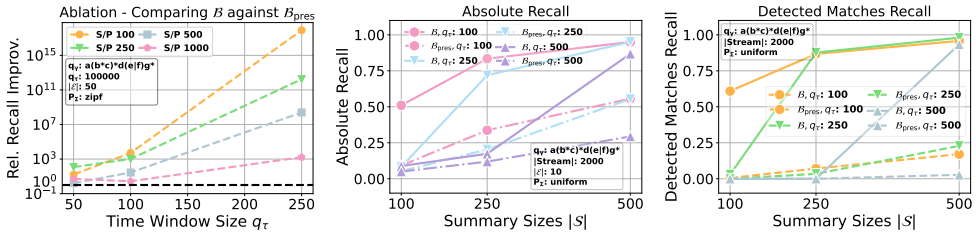


Figure 8: Examining the impact of the benefit function $\mathcal{B}$ compared to $\mathcal{B}_{\text{pres}}$ (a) on the relative recall improvement, (b) proximity to lower bound in absolute recall, and (c) proximity to lower bound in detected matches recall (higher is better in all plots).

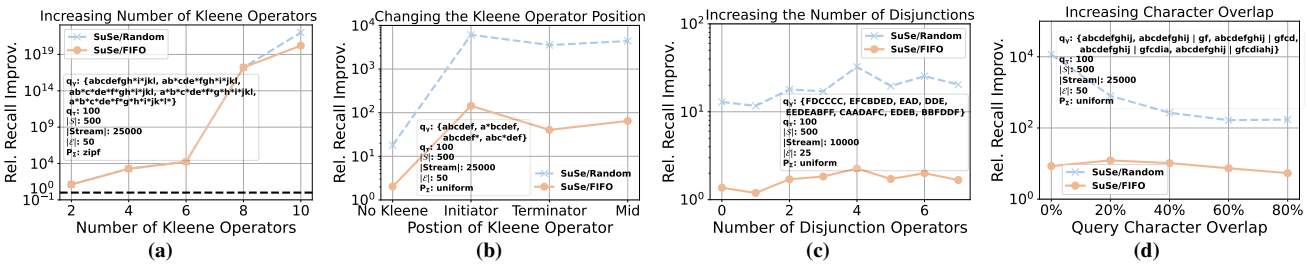


Figure 9: Examining the impact on relative recall improvement (higher is better) by varying (a) the count of Kleene stars, (b) Kleene star position, (c) disjunction operator count, and (d) query character overlap.

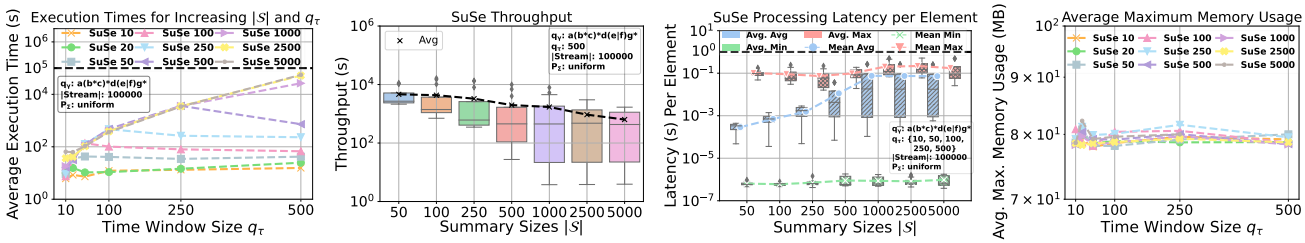
(100 to 5000), time window size q_τ (10 to 250), and alphabet probability distribution P_Σ (Zipfian, uniform, normal) affect the relative recall improvement. We adopt a query with $q_\gamma = a(b^*c)^*d(e|f)g^*$.

SUSE outperforms its baselines by choosing summaries that generate up to 10^{20} more matches. There is no parameter combination, in which SUSE performs worse than a baseline (black dashed line). A clear trend emerges for both plots: the larger q_τ and the smaller $|\mathcal{S}|$, the more significant the advantage. The larger the summary size, the more matches are found by chance by the baselines; nevertheless, for these parameters, SUSE chooses stream subsequences that generate at least three orders of magnitude more matches on average.

Ablation Study. We compared SUSE's benefit function $\mathcal{B}$ against a setup using solely the present benefit $\mathcal{B}_{\text{pres}}$ in terms of the relative recall improvement, the related absolute recall, and the detected matches recall. In Figure 8a, a trend similar to that in Figure 7a is observed: smaller $|\mathcal{S}|$ and larger q_τ lead to better subsequence selections. This is attributed to $\mathcal{B}$ providing a more accurate assessment of a character's future potential. For $q_\tau = 250$, SUSE with $\mathcal{B}$ obtains at least three orders of magnitude more matches on average over solely using $\mathcal{B}_{\text{pres}}$ and up to 10^{18} more for smaller values of $|\mathcal{S}|$, showing the effectiveness of $\mathcal{B}$.

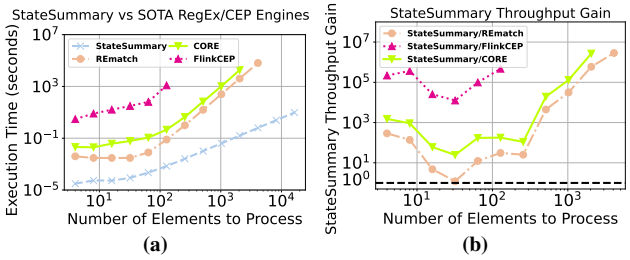
Absolute Recall. In Figure 8b, we investigated how close SUSE using $\mathcal{B}$ and $\mathcal{B}_{\text{pres}}$ comes to the optimal recall value. Increasing $|\mathcal{S}|$ positively affects recall for all methods. This improvement can be attributed to the reduced difference between the ground truth's summary size (equal to the stream size) and SUSE's summary size. Nonetheless, it is evident that SUSE, when employing expected benefits $\mathcal{B}$, attains higher recall values compared to using solely $\mathcal{B}_{\text{pres}}$. Also, when $|\mathcal{S}| \geq q_\tau$, with $\mathcal{B}$, SUSE selects substantially better subsequences, leading to a considerable increase in recall up to 80% – 95%. This trend can be attributed to our assumption in §6.2 regarding the expected benefit $\mathcal{B}$, where we implicitly assumed that at least one time window size of characters fits into the summary. For values $|\mathcal{S}| \geq q_\tau$, this assumption is valid, leading to more accurate estimates and an enhanced recall, boosting recall by approx. 70% compared to using only the present benefit $\mathcal{B}_{\text{pres}}$.

Detected Matches Recall. For Figure 8c, we examined how many of all possible matches were present in SUSE, again comparing $\mathcal{B}$ and $\mathcal{B}_{\text{pres}}$. The trends align closely with those observed for the absolute recall. SUSE with $\mathcal{B}$ yields a recall of 90% – 98% for $|\mathcal{S}| = 500$, indicating that SUSE chooses rich stream subsequences.



(a) (b) (c) (d)

Figure 10: (a) Execution time (lower is better) for processing 10^5 elements across different summary and time window sizes, (b) average minimum, average maximum, and average processing latency per element (lower is better).



(a) (b)

Figure 11: Comparison against the state of the art, focusing on (a) execution time (lower is better) and (b) throughput gain (higher is better) when processing varying numbers of elements.

7.3 Sensitivity Analysis

Next, we assess the effect of various parameters on the relative recall improvement. We fix $q_\tau = 100$ and $|\mathcal{S}| = 500$, and vary stream sizes (10000 - 25000), the numbers of evaluations (25 - 50), and P_Σ (Zipfian or uniform). We report averages of over 50 runs.

Number of Evaluation Timestamps. Figure 7c examines the influence of increasing the number of evaluation timestamps $|\mathcal{E}|$ on the relative recall improvement. As the boxplots indicate, increasing the number of evaluations does not impact the results.

Query Length. Figure 7d shows that as the query length grows, selecting subsequences that yield matches becomes increasingly challenging. SuSe addresses this by quantifying the importance of individual elements, prioritizing, for instance, rare elements that are essential to keep to obtain matches. While for query length two, FIFO performs slightly better, for query lengths above ten, SuSe is superior by at least one and up to almost five orders of magnitude.

Kleene Operator. In Figure 9a, we employed a RegEx of length twelve and incrementally increased the number of Kleene operators, distributing them randomly and uniformly. As SuSe is aware of characters with Kleene operators (by Count/Sum rules), it selects better summaries.

In Figure 9b, we examined four regular expressions: one without a Kleene operator and three with the Kleene's position varied. Without Kleene, SuSe's advantages are minimal, while there is a large difference for the other cases, independent of the Kleene position.

Disjunction Operator. We also explored the effect of the disjunction operator in Figure 9c. We varied the number of disjunctions from zero to seven and generated and combined random sequences for q_Y's with disjunction operators. Yet, the impact is negligible.

In Figure 9d, we fixed a query of length ten and progressively increased the character overlap. We randomly selected two characters of the given query for each step, linking them to the query and a disjunction. While the baseline show different trends, in all configurations, SuSe yields significant improvements.

7.4 Efficiency

State-of-the-art Comparison. We compared SuSe's efficiency against the state-of-the-art RegEx engine REmatch and two CEP engines, FlinkCEP and CORE, in Figure 11. Given the inferior performance of other RegEx engines (see [35]), we focus on REmatch as a representative RegEx engine. While FlinkCEP employs a traditional automata-based evaluation, CORE employs compact match representations, avoiding exponential state growth, and provides match enumeration with output-linear delay.

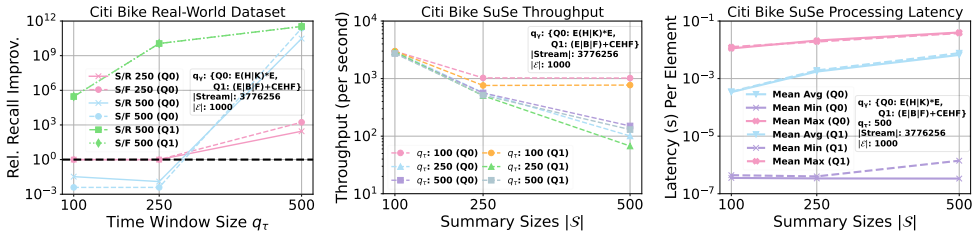
Both, the summary size and time window size are always set to the same x-value, so that the processed word fits within the summary entirely. Hence, all matches and aggregates can be computed by all methods, so that we focus exclusively on StateSummary operations (not employing summary selection). We tested with $q_Y = ABCD$ (the REmatch syntax being $!x\{A\}.*!y\{B\}.*!z\{C\}.*!w\{D\}$) for the required processing time of input words of varying sizes. As input, we choose words of the form $A^iB^iC^iD^i$ with $i \in \mathbb{N}^+$, e.g., $A^8B^8C^8D^8$, ensuring an increase in matches for rising x-values.

SuSe's processing time is consistently at least an order of magnitude less than REmatch and CORE, and at least four orders of magnitude less than Flink, due to the StateSummary. Particularly for words longer than 128 characters, the performance gap between SuSe and both REmatch and CORE significantly widens, while Flink did not terminate for these values due to overload.

The performance differences also induce throughput gains, as shown in Figure 11b. The ratio indicates how much more quickly SuSe processed a word than REmatch, CORE, or Flink. The throughput gain is always greater than one, indicating the superiority of SuSe. For longer words, this gain increases significantly.

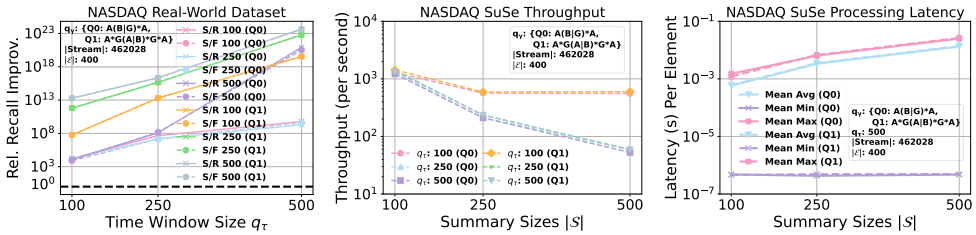
Overall Efficiency. In Figure 10, we examined how varying $|\mathcal{S}|$ (see legend) and q_τ affects the execution time, throughput, processing latency per element, and memory usage. Figure 10a illustrates that the larger $|\mathcal{S}|$ and the larger q_τ, the longer the execution time.

Also, a time window size exists for all procedures up to a summary size of 500, so that the execution time no longer increases but slightly decreases. We explain this as follows: as shown in §5.2, the Remove operation is the most expensive. Its cost is dictated by $min(q_\tau, |\mathcal{S}|)$. Once either $|\mathcal{S}| > q_\tau$ or $q_\tau > |\mathcal{S}|$, the smaller value of the two parameters becomes the limiting factor for runtime, rendering the influence of the other parameter's increase marginal. With larger q_τ,



(a) (b) (c)

Figure 12: Citi Bike: (a) relative recall improvement (higher is better), (b) throughput (higher is better), and (c) latency (lower is better).



(a) (b) (c)

Figure 13: NASDAQ: (a) relative recall improvement (higher is better), (b) throughput (higher is better), and (c) latency (lower is better).

SUSE chooses better subsequences, which reduces the number of REMOVE operations and explains the drop in runtime for $q_\tau \geq |\mathcal{S}|$.

In Figure 10b, we examined the resulting throughput values for a fixed $q_\tau = 500$. For the same reasons as discussed above, there is a decline in throughput as the summary size expands.

In Figure 10c, we investigated the resulting average min, max, and average processing latency per element per second. A boxplot contains the latencies that arose over increasing q_τ, while the corresponding line plots represent the averages. A larger $|\mathcal{S}|$ leads to elevated processing latencies. However, the median of the average processing latencies consistently remains below $10^{-1}s$.

Finally, we examined the induced memory footprint depicted in Figure 10d. Neither increasing $|\mathcal{S}|$ nor q_τ influences average maximum memory usage. With our fixed size of 128 bits for counters, the impact on memory from increasing $|\mathcal{S}|$ is negligible.

7.5 Case Study: Citi Bike

Character Types and Queries. Figure 12 denotes our results for the Citi Bike real-world dataset. The character types encompass rides on frequented routes, pinpoint brief rides at busy stations, rides at central stations, and member rides at quieter stations, giving a thorough understanding of the Citi Bike dynamics. We defined a query with $Q0_\gamma = E(H|K)^*E$, recognizing patterns where a ride starts at a busy station, followed by a combination of rides at central and quieter stations, ending with a ride at a busy station, suggesting probable areas for station enhancements or upkeep. The second query, $Q1_\gamma = (E|B|F)^+CEHF$, captures sequences of rides that start at busy stations, popular routes, or extended rides at central spots, followed by short and consecutive rides at busy stations, a trip at a central location, and concluding with a prolonged ride there.

Effectiveness. To examine the relative recall improvement, we varied $|\mathcal{S}|$ and q_τ in Figure 12a. For both queries, not all summary sizes yielded matches. Generally, there is an upward trend in the results with a larger q_τ; SUSE identifies subsequences that are between three and twelve orders of magnitude superior for $q_\tau = 500$.

Throughput and Latency. We analyzed the throughput and processing latency of SUSE in Figure 12b and Figure 12c. For smaller values of $|\mathcal{S}|$ and q_τ (refer to legend), SUSE's throughput increases. However, with larger values for $|\mathcal{S}|$ and q_τ, the lower throughput is attributed to the REMOVE operation and memory shift overheads.

Figure 12c shows the avg min, avg max, and avg processing latency of SUSE, with $q_\tau = 500$. An increase in summary size leads to higher latencies. Yet, the average max latency hovers around 10^{-2} seconds, while the average processing latency remains under 10^{-2} seconds for $|\mathcal{S}| = 500$, indicating real-time computing latencies.

7.6 Case Study: NASDAQ

Character Types and Queries. Figure 13 represents our results for the NASDAQ real-world dataset. We derived character types based on stock market activities, e.g., significant price changes, trade volumes, daily peak or lowest prices, periods of consistent behaviour or fluctuations, and market uncertainty. We formulated a query with $Q0_\gamma = A(B|G)^*A$, detecting an initial price increase, succeeded by any number of price reductions or uncertain market periods, and ending with a subsequent price surge, potentially signalling a fluctuating stock price ascent. The second query, $Q1_\gamma = A^*G(A|B)^*G^*A$, captures zero or more price rises, followed by market uncertainty, zero or more occurrences of either significant price rises or drops, a potential period of market uncertainty, and ends with a price rise, indicating a trend that, despite its volatility and periods of uncertainty, ends with a bullish behaviour.

Effectiveness. In Figure 13a, we varied $|\mathcal{S}|$ and q_τ between $100-500$ using the $Q0$ and $Q1$ above. Here, for all $|\mathcal{S}|$, matches were obtained. Generally, larger $|\mathcal{S}|$ and q_τ led SuSe to derive superior stream projections, resulting in up to 10^{23} additional matches. However, even at smaller parameter values, SuSe's projections surpassed both baselines by magnitudes ranging from 10^4 to 10^{13}. Notably, for $Q1$, SuSe identified better projections due to the increased count of Kleene operators.

Throughput and Latency. In our examination of throughput and processing latency of SuSe, as shown in Figure 13b and Figure 13c, a similar trend as for Citi Bike emerges: when decreasing the summary and time-window size (see legend), SuSe's throughput enhances, reaching peaks of up to 1500 elements per second. However, when the $|\mathcal{S}|$ and q_τ increase, the throughput decreases to around $70-80$ elements.

Turning to Figure 13c, we fixed $q_\tau = 500$ and increased the summary size, resulting in an upward trend of the latencies. Nevertheless, while the average maximum latency approaches 10^{-2} seconds, the average processing latency for $\mathcal{S} = 500$ stays just below 10^{-2} seconds, again denoting latencies suitable for real-time computing.

8 RELATED WORK

Subsequences. Subsequences are crucial tools in fields like biological sequence analysis [33] and event stream processing [4, 14, 43]. They are associated with algorithmic challenges including the longest common subsequence [5], shortest common supersequence [29], or subsequence counting [11], and are studied in contexts of gap-size [10] and wildcard constraints [20, 21]. However, obtaining aggregated information about subsequence *mappings* within a text has barely been researched. Solely in [11] and [15], dynamic programming algorithms are presented which for a given subsequence u and a text x count the number of mappings $u \preceq_m x$. We complement their work using a more holistic approach: We provide aggregated information for a set of patterns derived from a regular expression γ and simultaneously account for partial mappings from $L_{\text{pre}}(\gamma)$. While prior works assume static text, SuSe processes (valued) character streams, which require online computation.

CEP Optimizations. CEP facilitates reactive programming by analyzing queries over event streams. CEP queries primarily utilize regular expressions, enabling pattern recognition against incoming events. Various strategies address system overload, including load shedding techniques [7, 37, 44]. These techniques selectively discard events or remove partial matches unlikely to complete, reducing information loss while decreasing overload. Similarly, filtering methods [3] transform a stream s into a filtered stream s', retaining events more likely to lead to complete matches and omitting others. SuSe also functions as a filter, with its selected subsequence representing stream s'. It resembles these methods in (1) state-based decisions, (2) reducing system load, and (3) minimizing information loss. Yet, SuSe bases decisions on the StateSummary without materializing the state, evading the exponential complexity of automata-based evaluation, and its decisions, unlike [3, 44], are not reliant on a learned model, enhancing streaming practicality.

9 CONCLUSIONS

We proposed SuSe, an architecture for regular expression subsequence summarization over data streams. It incorporates a StateSummary data structure that captures a query-specific stream summary through aggregated subsequence match information. We presented a summary selection algorithm, leveraging StateSummary for choosing stream subsequences, which aim to minimize the loss in the aggregated results over time. Our evaluation on real-world and synthetic data validates the efficiency and effectiveness of SuSe: it processes inputs by multiple orders of magnitude faster than leading RegEx and CEP engines while generating aggregates based on substantially richer stream subsequences than baseline approaches.

REFERENCES

[1] Oniguruma – a modern and flexible regular expressions library. 2022. https://github.com/kkos/oniguruma Accessed on 2023-10-15.

[2] Jagrati Agrawal, Yanlei Diao, Daniel Gyllstrom, and Neil Immerman. 2008. Efficient pattern matching over event streams. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, Jason Tsong-Li Wang (Ed.). ACM, 147–160. https://doi.org/10.1145/1376616.1376634

[3] Adar Amir, Ilya Kolchinsky, and Assaf Schuster. 2022. DLACEP: A Deep-Learning Based Framework for Approximate Complex Event Processing. In *Proceedings of the 2022 International Conference on Management of Data*. ACM. https://doi.org/10.1145/3514221.3526136

[4] Alexander Artikis, Alessandro Margara, Martín Ugarte, Stijn Vansummeren, and Matthias Weidlich. 2017. Complex Event Recognition Languages: Tutorial. In *Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems, DEBS 2017, Barcelona, Spain, June 19-23, 2017*. ACM, 7–10. https://doi.org/10.1145/3093742.3095106

[5] Ricardo A. Baeza-Yates. 1991. Searching subsequences. *Theoretical Computer Science* 78, 2 (1991), 363–376. https://doi.org/10.1016/0304-3975(91)90358-9

[6] Marco Bucchi, Alejandro Grez, Andrés Quintana, Cristian Riveros, and Stijn Vansummeren. 2022. CORE: a complex event recognition engine. *Proceedings of the VLDB Endowment* 15, 9 (May 2022), 1951–1964. https://doi.org/10.14778/3538598.3538615

[7] Koral Chapnik, Ilya Kolchinsky, and Assaf Schuster. 2021. DARLING: Data-Aware Load Shedding in Complex Event Processing Systems. *Proc. VLDB Endow.* 15, 3 (2021), 541–554. http://www.vldb.org/pvldb/vol15/p541-chapnik.pdf

[8] Hao Chen, Yu Chen, and Douglas H. Summerville. 2011. A Survey on the Application of FPGAs for Network Infrastructure Security. *IEEE Communications Surveys & Tutorials* 13, 4 (2011), 541–561. https://doi.org/10.1109/surv.2011.072210.00075

[9] citi Bike. 2022. http://www.citibikenyc.com/system-data..

[10] Joel D. Day, Maria Kosche, Florin Manea, and Markus L. Schmid. 2022. Subsequences with Gap Constraints: Complexity Bounds for Matching and Analysis Problems. In *33rd International Symposium on Algorithms and Computation (ISAAC 2022) (LIPIcs)*, Sang Won Bae and Heejin Park (Eds.), Vol. 248. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 64:1–64:18. https://doi.org/10.4230/LIPIcs.ISAAC.2022.64

[11] Cees Elzinga, Sven Rahmann, and Hui Wang. 2008. Algorithms for subsequence combinatorics. *Theoretical Computer Science* 409, 3 (Dec. 2008), 394–404. https://doi.org/10.1016/j.tcs.2008.08.035

[12] PCRE2 – Perl-Compatible Regular Expressions. 2022. https://github.com/PCRE2Project/pcre2 Accessed on 2023-10-15.

[13] Apache Software Foundation. 2021. Apache Flink. https://nightlies.apache.org/flink/flink-docs-release-1.14/ Accessed on 2023-10-15.

[14] Nikos Giatrakos, Elias Alevizos, Alexander Artikis, Antonios Deligiannakis, and Minos N. Garofalakis. 2020. Complex event recognition in the Big Data era: a survey. *VLDB J.* 29, 1 (2020), 313–352. https://doi.org/10.1007/s00778-019-00557-w

[15] Ronald I. Greenberg. 2003. Computing the Number of Longest Common Subsequences. arXiv:arXiv:cs/0301034

[16] PCREgrep – A grep program that uses the PCRE regular expression library. 2014. https://github.com/vmg/pcre/blob/master/pcregrep.c/ Accessed on 2023-10-15.

[17] John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. 2001. Introduction to automata theory, languages, and computation, 2nd edition. *ACM SIGACT News* 32, 1 (March 2001), 60–65. https://doi.org/10.1145/568438.568455

[18] Theodore Johnson, S. Muthukrishnan, and Irina Rozenbaum. 2007. Monitoring Regular Expressions on Out-of-Order Streams. In *2007 IEEE 23rd International Conference on Data Engineering*. IEEE. https://doi.org/10.1109/icde.2007.369001

[19] Stephen C. Kleene. 1951. *Representation of events in nerve nets and finite automata*. Technical Report. RAND Corporation, Santa Monica, CA.

[20] Sarah Kleest-Meißner, Rebecca Sattler, Markus L. Schmid, Nicole Schweikardt, and Matthias Weidlich. 2022. Discovering Event Queries from Traces: Laying Foundations for Subsequence-Queries with Wildcards and Gap-Size Constraints. In *25th International Conference on Database Theory, ICDT 2022 (LIPIcs)*, Dan Olteanu and Nils Vortmeier (Eds.), Vol. 220. Schloss Dagstuhl, 18:1–18:21. https://doi.org/10.4230/LIPIcs.ICDT.2022.18

[21] Sarah Kleest-Meißner, Rebecca Sattler, Markus L. Schmid, Nicole Schweikardt, and Matthias Weidlich. 2023. Discovering Multi-Dimensional Subsequence Queries from Traces - From Theory to Practice. In *Datenbanksysteme für Business, Technologie und Web (BTW 2023), Proceedings (LNI)*, Birgitta König-Ries, Stefanie Scherzinger, Wolfgang Lehner, and Gottfried Vossen (Eds.), Vol. P-331. GI e.V., 511–533. https://doi.org/10.18420/BTW2023-24

[22] Sailesh Kumar, Balakrishnan Chandrasekaran, Jonathan Turner, and George Varghese. 2007. Curing regular expressions matching algorithms from insomnia, amnesia, and acalculia. In *Proceedings of the 3rd ACM/IEEE Symposium on Architecture for networking and communications systems*. ACM. https://doi.org/10.1145/1323548.1323574

[23] Sailesh Kumar, Sarang Dharmapurikar, Fang Yu, Patrick Crowley, and Jonathan Turner. 2006. Algorithms to accelerate multiple regular expressions matching for deep packet inspection. In *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*. ACM. https://doi.org/10.1145/1159913.1159952

[24] Boost Regex Library. 2022. https://github.com/boostorg/regex Accessed on 2023-10-15.

[25] NASDAQ Data Link. 2023. https://data.nasdaq.com/ Accessed: 2023-10-15.

[26] Alex X. Liu and Eric Norige. 2019. A De-Compositional Approach to Regular Expression Matching for Network Security. *IEEE/ACM Transactions on Networking* 27, 6 (Dec. 2019), 2179–2191. https://doi.org/10.1109/tnet.2019.2941920

[27] Tingwen Liu, Yong Sun, Alex X. Liu, Li Guo, and Binxing Fang. 2012. A Prefiltering Approach to Regular Expression Matching for Network Security Systems. In *Applied Cryptography and Network Security*. Springer Berlin Heidelberg, 363–380. https://doi.org/10.1007/978-3-642-31284-7_22

[28] Lei Ma, Chuan Lei, Olga Poppe, and Elke A. Rundensteiner. 2022. Gloria: Graph-based Sharing Optimizer for Event Trend Aggregation. In *Proceedings of the 2022 International Conference on Management of Data*. ACM. https://doi.org/10.1145/3514221.3526145

[29] David Maier. 1978. The Complexity of Some Problems on Subsequences and Supersequences. *Journal of the ACM* 25, 2 (April 1978), 322–336. https://doi.org/10.1145/322063.322075

[30] Olga Poppe, Chuan Lei, Lei Ma, Allison Rozet, and Elke A. Rundensteiner. 2021. To Share, or not to Share Online Event Trend Aggregation Over Bursty Event Streams. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, Guoliang Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava (Eds.). ACM, 1452–1464. https://doi.org/10.1145/3448016.3452785

[31] Olga Poppe, Chuan Lei, Elke A. Rundensteiner, and David Maier. 2017. GRETA: graph-based real-time event trend aggregation. *Proceedings of the VLDB Endowment* 11, 1 (Sept. 2017), 80–92. https://doi.org/10.14778/3151113.3151120

[32] Olga Poppe, Chuan Lei, Elke A. Rundensteiner, and David Maier. 2019. Event Trend Aggregation Under Rich Event Matching Semantics. In *Proceedings of the 2019 International Conference on Management of Data*. ACM. https://doi.org/10.1145/3299869.3319862

[33] Sven Rahmann. 2006. Subsequence Combinatorics and Applications to Microarray Production, DNA Sequencing and Chaining Algorithms. In *Combinatorial Pattern Matching*. Springer Berlin Heidelberg, 153–164. https://doi.org/10.1007/11780441_15

[34] RE2 regular expression library. 2022. https://github.com/google/re2 Accessed on 2023-10-15.

[35] Cristian Riveros, Nicolás Van Sint Jan, and Domagoj Vrgoč. 2023. REmatch: A Novel Regex Engine for Finding All Matches. *Proceedings of the VLDB Endowment* 16, 11 (July 2023), 2792–2804. https://doi.org/10.14778/3611479.3611488

[36] Allison Rozet, Olga Poppe, Chuan Lei, and Elke A. Rundensteiner. 2020. Muse: Multi-query Event Trend Aggregation. In *Proceedings of the 29th ACM International Conference on Information &amp Knowledge Management*. ACM. https://doi.org/10.1145/3340531.3412138

[37] Ahmad Slo, Sukanya Bhowmik, and Kurt Rothermel. 2020. hSPICE: state-aware event shedding in complex event processing. In *Proceedings of the 14th ACM International Conference on Distributed and Event-based Systems*. ACM. https://doi.org/10.1145/3401025.3401742

[38] Peter Snyder and Chris Kanich. 2015. No Please, After You: Detecting Fraud in Affiliate Marketing Networks.. In *WEIS*.

[39] Steven Purtzel and Matthias Weidlich. 2024. SuSe: Summary Selection for Regular Expression Subsequence Aggregation over Streams. https://github.com/spurtzel/SuSe.

[40] Ken Thompson. 1968. Programming Techniques: Regular expression search algorithm. *Commun. ACM* 11, 6 (June 1968), 419–422. https://doi.org/10.1145/363347.363387

[41] robust TRE – a lightweight and efficient POSIX compliant regexp matching library. 2021. https://github.com/laurikari/tre Accessed on 2023-10-15.

[42] Mohamed Zaki and Babis Theodoulidis. 2013. Analyzing Financial Fraud Cases Using a Linguistics-Based Text Mining Approach. *SSRN Electronic Journal* (2013). https://doi.org/10.2139/ssrn.2353834

[43] Haopeng Zhang, Yanlei Diao, and Neil Immerman. 2014. On complexity and optimization of expensive queries in complex event processing. In *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*, Curtis E. Dyreson, Feifei Li, and M. Tamer Özsu (Eds.). ACM, 217–228. https://doi.org/10.1145/2588555.2593671

[44] Bo Zhao, Nguyen Quoc Viet Hung, and Matthias Weidlich. 2020. Load Shedding for Complex Event Processing: Input-based and State-based Techniques. In *36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20-24, 2020*. IEEE, 1093–1104. https://doi.org/10.1109/ICDE48307.2020.00099