Space-Economical Partial Gram Indices for Exact Substring Matching

Nan Tang
CWI, Amsterdam
The Netherlands
tang@cwi.nl

Lefteris Sidirourgos
CWI, Amsterdam
The Netherlands
lsidir@cwi.nl

Peter Boncz
CWI, Amsterdam
The Netherlands
boncz@cwi.nl

ABSTRACT

Exact substring matching queries on large data collections can be answered using \( q \)-gram indices, that store for each occurring \( q \)-byte pattern an (ordered) posting list with the positions of all occurrences. Such gram indices are known to provide fast query response time and to allow the index to be created quickly even on huge disk-based datasets. Their main drawback is relatively large storage space, that is a constant multiple (typically \( > 2 \)) of the original data size, even when compression is used. In this work, we study methods to conserve the scalable creation time and efficient exact substring query properties of gram indices, while reducing storage space. To this end, we first propose a partial gram index based on a reduction from the problem of omitting indexed \( q \)-grams to the set cover problem. While this method is successful in reducing the size of the index, it generates false positives at query time, reducing efficiency. We then increase the accuracy of partial grams by splitting posting lists of frequent grams in a frequency-tuned set of signatures that take the bytes surrounding the grams into account. The resulting \( qs \)-gram scheme is tested on huge collections (up to 426GB) and is shown to achieve an almost 1:1 data:index size, and query performance even faster than normal gram methods, thanks to the reduced size and access cost.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: Miscellaneous

General Terms

Algorithms, Design, Performance

1. INTRODUCTION

Finding all instances of an ordered sequence of bytes (i.e., a string) in a large file is a fundamental pattern matching problem. Efficient solutions with real-time performance are attractive to many applications, e.g., online string search, word processing software, computational biology and digital forensic pattern search. Both the database and IR communities have investigated many string search topics, for instance approximate (e.g., edit distance based) string selection and top-K joins [1, 10, 12, 20], full text search [2, 13], and digital forensic search [18]. In this work we stick to the basic problem of exact substring matching [14, 16], but note that techniques for solving this are typically the building blocks to address elaborate approximate string matching problems.

We focus on the following efficiency aspects of the exact substring search indexing problem:

- index creation time,
- index space, and
- query response time.

These aspects are influenced by the following dimensions:

- data size (scalability),
- data distribution, in particular the frequency distribution of co-located letters, and
- the length of the queries.\(^1\)

Ideally, one would like to achieve an index that takes similar space (or less) as the input data, can be created in time close to a sequential I/O pass over that data, and provides query response time close to a sequential I/O pass over the query result volume; with any data and query distribution. These objectives are very hard to achieve all together, and get even harder when considering datasets with a uniform rather than a skewed distribution.

**Suffix Techniques.** In the algorithms community, the suffix tree [8, 16] and suffix array [9, 14] are being studied as indices to allow fast substring searching. Due to the large amount of information carried in each node and edge of a suffix tree, the storage overhead is 10-20x the input data in good implementations, which renders suffix trees impractical in most applications. The suffix array is a more space efficient variant which simply constructs an array of positions, where the positions point to the suffixes in the string in lexicographical order. For both suffix trees as well as

\(^1\)We assume a distribution of co-located query letters similar to the data distribution.
suffix arrays, algorithms are known whose space and time complexity are \(O(|\sigma|)\), where \(|\sigma|\) is the length of input data string \(\sigma\). The excellent property of suffix arrays is that regardless of the query (long or short, frequent or rare), lookup is simply a binary search of fixed cost. Sometimes this binary search, which leads to \(\log|\sigma|\) (on string \(\sigma\)) I/Os, is held against suffix arrays in favour of potentially \(O(1)\) hash-based solutions. However, this disadvantage can be mitigated by creating a sparse RAM resident B-tree index on top of the array. Regrettably, however, all known suffix array construction algorithms do not truly scale to large datasets. Once the data size exceeds main memory, index creation performance strongly deteriorates due to the need for close to \(|\sigma|\) random disk writes. As a result, the best scaling reported “linear” suffix array construction algorithm (using special Linux kernel patches to improve random I/O) has been demonstrated on only 5GB of data [9]. Even on this effectively small – potentially RAM resident – data size, suffix array construction takes many hours. Our work focuses on truly scalable techniques that can be used to manage huge disk-based datasets, where the input data set can – in the case of forensic data search – be a set of full hard drive images (terabyte scale and beyond). As the scalability aspect of suffix methods is several orders of magnitude off target for our objectives, we focus on alternative methods.

Gram Techniques. The basic idea of q-grams is to construct an index on all occurring patterns of q co-located letters (we stick to bytes, in this work – depending on the alphabet a byte can either contain more or less than one letter). Figure 1 shows a string for a slogan of the Beijing 2008 Olympic games. Generally speaking, the whole string is indexed using q-grams, with each q-gram associated with a list of postings e.g. 3-gram postings as shown in Figure 1. Postings can be stored using IR engines specialised in inverted list processing, but it has been shown that a DBMS can also be used to store and query such data efficiently [10]. A given query \(q\) (for exact match) is also decomposed to several q-grams \(G\). By performing merge join on the posting lists associated with each q-gram in \(G\), we can identify all occurrences of \(q\). For example, to find a phrase with the first word ‘one’ and the second word starting with ‘w’, we can submit a 5-byte query ‘one_w’. By doing merge join of the three posting lists of ‘one’ (1, 11, 21), ‘ne_w’ (2, 12, 22) and ‘e_w’ (3), we could identify one match at posting 1. IR systems routinely optimize posting list processing by choosing a merge join that puts the shortest lists first. Also, it is not strictly required to join the posting lists of all occurring grams; posting lists may be pruned (omitted from the plan) as long as all letters in the query remain covered by at least one non-pruned gram. A typical approach in gram query processing is thus to order the shortest posting lists first in the plan and prune the longest lists as long as the query remains covered.

Comparison. The creation of a gram-based (inverted) in-

Figure 1: A String and Its Posting Lists

<table>
<thead>
<tr>
<th>Dataset</th>
<th>3-gram</th>
<th>4-gram</th>
<th>5-gram</th>
<th>6-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td>wikipedia 2.1GB</td>
<td>5.4x10^6</td>
<td>1.4x10^6</td>
<td>1.8x10^6</td>
<td>2.5x10^6</td>
</tr>
<tr>
<td>INDEX 1.5GB</td>
<td>5.4x10^6</td>
<td>1.4x10^6</td>
<td>1.8x10^6</td>
<td>2.5x10^6</td>
</tr>
<tr>
<td>acquaint 3.0GB</td>
<td>1.8x10^6</td>
<td>1.4x10^6</td>
<td>7.9x10^6</td>
<td>2.2x10^6</td>
</tr>
<tr>
<td>XMLmark 3.4GB</td>
<td>4.9x10^6</td>
<td>4.0x10^6</td>
<td>1.8x10^6</td>
<td>5.8x10^6</td>
</tr>
<tr>
<td>movielens 690MB</td>
<td>4.2x10^6</td>
<td>7.8x10^6</td>
<td>8.9x10^6</td>
<td>9.0x10^6</td>
</tr>
<tr>
<td>moviemebb 1.2GB</td>
<td>4.2x10^6</td>
<td>1.3x10^6</td>
<td>1.5x10^6</td>
<td>1.5x10^6</td>
</tr>
</tbody>
</table>

Table 2: q-gram statistics for various datasets
showing that the \( q \)-gram approach combines good query performance with strongly reduced index space. After discussing related work in Section 6, we conclude in Section 7.

2. PRELIMINARIES

A string \( \sigma \) is defined to be an ordered sequence of bytes, and its length to be \(|\sigma| = n\). We denote \( \sigma[i, j], 1 \leq i \leq j \leq n \) the substring of \( \sigma \) of length \( j - i + 1 \) starting at position \( i \). If \( i = j \) then we write \( \sigma[i] \).

A \( q \)-gram is a string of length \( q \). A \( q \)-gram \( g \) appears at position \( i \) of the string \( \sigma \), if \( g = \sigma[i, i + q - 1] \). The set of all \( q \)-grams that appear in a string \( \sigma \) is denoted by \( G = \{g_1, \ldots, g_k\} \). Each \( q \)-gram \( g_i \in G \) is associated with a list of postings \( P(g_i) = \{p_1, \ldots, p_j\} \), where each posting \( p_j \in \{1, \ldots, n - q + 1\} \) refers to \( i \) \in \{1, \ldots, l\} \) to a position in the string \( \sigma \) where \( g_i \) appears. Evidently, a \( q \)-gram \( g \) can appear more than once in a string \( \sigma \).

A \( q \)-gram \( g \) covers the bytes \( \sigma[i], \ldots, \sigma[i + q - 1] \) if there is a posting \( p \in P(g) \) such that \( p = i \). A set \( G \) of \( q \)-grams covers the entire string \( \sigma \), if every byte \( \sigma[i] \) of the string \( \sigma \) is covered by at least one \( q \)-gram.

Finally, a \emph{string matching query} is defined to be the query which given the string \( \sigma \) of length \( |\sigma| \), requests all positions \( i \) of the string \( \sigma \) such that \( g = \sigma[i, i + |g| - 1] \).

Q-GRAM based indices utilise the set of \( q \)-grams \( G \) to efficiently answer queries on approximate and exact substring match, text auto-completion, and for error-correction. The general approach, which will be referred to in the sequel as the \emph{full \( q \)-gram index (FG)}, covers each byte of the string \( \sigma \) exactly \( q \) times. This is achieved by including in the set \( G \) all possible \( q \)-grams of the string \( \sigma \), such that each byte \( \sigma[i] \) may be covered by those \( q \)-grams that have postings equal to \( i - q + 1, \ldots, i \). The most common choice for \( q \) is 3. The total number of postings in a full \( q \)-gram index is \(|\sigma| - q + 1 \).

In this work, we argue that for \emph{space-economical} \( q \)-gram based indices it is better to use \emph{partial \( q \)-gram indices (PG)} instead of full. However, for providing complete answers to substring matching queries, we have to ensure that each byte in the string \( \sigma \) is covered at least once. The problem at hand is formally defined as follows: \emph{Given a string \( \sigma \) and its set \( G \) of \( q \)-grams, find a minimal subset \( G' \subset G \), such that \( G' \) covers the entire string \( \sigma \).} The coverage requirement is necessary to ensure that no matching substrings are missed, thus resulting in false negatives. However, a partial \( q \)-gram index will return exact matches plus a small list of candidate substrings because only a subset of \( G \) is indexed. In Section 5 we experimentally compare the number of candidates and the actual matches for the partial \( q \)-gram indices.

Another important difference between a full \( q \)-gram index and a partial one is that the former can match any substring that is no shorter than \( q \), while the latter loses some expressive power: it can find matching candidates for a substring that is no shorter than \( 2q - 1 \).

**Lemma 1.** If \( G' \) is a partial set of \( q \)-grams for string \( \sigma \) and \( g \) is a query on \( \sigma \) that can be evaluated correctly, then the length \(|g| \) of \( g \) is no shorter than \( 2q - 1 \).

**Proof.** There are 3 cases. First, if \(|g| < q \) then it is impossible to match any \( q \)-gram with the query \( g \) since the length of \( g \) is smaller than the length of the indexed \( q \)-grams\(^2\). The second case is when \(|g| \leq |q| = 2q - 2 \).

\(^2\)We discount the non-practical approach of merging all superset \( P_q \{P_x \in G': x \subset q\} \) into a list of candidates.

Let query \( q \) match on string \( \sigma \) at positions \( \sigma[i, i + 2q - 3] \). Also, assume that these bytes of \( \sigma \) are covered by the two \( q \)-grams \( g_1, g_2 \) with postings \( P(q_1) = \sigma[i - 1, i + q - 1] \) and \( P(q_2) = \sigma[i + q, i + 2q - 2] \). In this case, no \( q \)-gram that can be extracted from query \( q \) will match any \( q \)-gram used to cover the bytes of string \( \sigma \) from position \( i \) to \( i + 2q - 3 \), and thus false negatives will appear. Finally, the third case is when \(|g| \geq 2q - 1 \), then there should always be a \( q \)-gram that appears on both the query \( q \) and the string \( \sigma \). If no such \( q \)-gram exists, then the query result is empty.

**Positional Merge Join.** Consider two \( q \)-grams \( g_1, g_2 \) and their respective postings \( P(g_1) = \{p_1, \ldots, p_k\} \) and \( P(g_2) = \{p'_1, \ldots, p'_l\} \) over string \( \sigma \). Also, consider a query \( q \) where the \( q \)-gram \( g_1 \) appears in position \( i \) of \( q \), and \( q \)-gram \( g_2 \) in position \( i + k \) of \( q \), where \( i + k < \) \(|q| = 2q - 1 \). In order to find which postings of \( g_1 \) and \( g_2 \) are valid candidates for matching query \( q \) to string \( \sigma \), we have to join \( P(g_1) \) with \( P(g_2) \). This join must take into consideration also the distance \( k \) between \( g_1 \) and \( g_2 \) in order to find the same distance must be preserved in \( \sigma \), too. The positional merge join is the join operation of two posting lists and an offset \( k \).

For the two posting lists \( P(g_1), P(g_2) \) and an offset \( k \), the operation \emph{PosMergeJoin}(\( P(g_1), P(g_2), k \)) is defined as:

\[
\begin{align*}
\text{SELECT} & \ P(g_1).\text{posting} \\
\text{FROM} & \ P(g_1).\text{posting}, P(g_2).\text{posting} \\
\text{WHERE} & \ P(g_1).\text{posting} = P(g_2).\text{posting} - k
\end{align*}
\]

**Example 2.1.** Let assume that there is a partial \( 3 \)-gram index for the string in Figure 1. This partial index includes the \( 3 \)-gram ‘one’ with postings list \( P(‘one’) = \{1, 11, 21\} \) and the \( 3 \)-gram ‘e’ \( \cdot \) ‘w’ with postings list \( P(‘e’ \cdot ‘w’) = \{3\} \). Let the query \( q = ‘one’ \cdot ‘w’ \). Query \( q \) is decomposed to three \( 3 \)-grams ‘one’, ‘e’ \( \cdot \) ‘w’, and ‘e’ \( \cdot \) ‘w’. Since only \( 3 \)-gram ‘one’ and ‘e’ \( \cdot \) ‘w’ exist in the partial index, we apply the positional merge join operator on the postings lists of these two \( 3 \)-grams with an offset \( 2 \). The result is posting \( \{1\} \), which is the position that \( q \) matches the string in Figure 1.

The positional merge join will return a candidate list of positions in the string \( \sigma \) for the query \( q \) to be verified. This happens when there is no \( q \)-gram in the index that covers the first and last \( q-1 \) bytes of the query \( q \). However, all bytes in between should be matched with some \( q \)-grams in the partial index. If this is not the case, then it is safe to conclude that there is no occurrences of query \( q \) in \( \sigma \). This provides opportunity during query evaluation, to reject mismatches early, and to only verify candidates in the raw data (i.e., the string \( \sigma \)) at the very end, only for those queries whose borders were not fully covered.

3. PARTIAL Q-GRAM INDICES

This section presents techniques for choosing the appropriate set of \( q \)-grams to be included in a partial \( q \)-gram index. To this end, we employ existing techniques from the set-theory field and extend them to meet practical requirements in terms of indexing and querying time. In addition, we introduce \emph{signature} and \emph{hash based} \( q \)-grams and then propose a novel \( q \)-gram approach.

**Partial Q-Gram Selection.** The selection of the \( q \)-grams to be included in the partial index should satisfy the following two objectives: i) each \( q \)-gram must have a sorted list of postings in order to minimise I/O during posting fetching,
input: σ = the input string. k = the length of the grams
output: a set of k-grams G whose postings cover σ

1. G = ∅;
2. scan σ and derive all posting lists \( P(g_i), g_i \in G \);
3. calculate \( F(g_i) = |P(g_i)| \) for all \( g_i \in G \);
4. sort all \( F(g_i) \) in descending order;
5. cnt(\( g_i \)) = 0 for \( i \in \{1, \ldots, |\sigma|\} \);
6. for each \( g_i \) in descending order of \( F(g_i) \) do
7. selected \( = \) false;
8. if \( q \) is the first or the last \( q \)-gram in \( \sigma \) then
9. \( G' = G' \cup \{g_i\} \);
10. selected \( = \) true;
11. else
12. for each posting \( p \) in \( P(g_i) \) and selected \( = \) false do
13. for \( k \) from 0 to \( q - 1 \) do
14. if cnt(p) \( + k \) = \( q - 1 \) then
15. \( G' = G' \cup \{g_i\} \);
16. break;
17. end for each posting \( p \) in \( P(g_i) \) do
18. for \( k \) from 0 to \( q - 1 \) do
19. if selected then
20. cnt(p) \( + k \) \( = \) q;
21. else if cnt(p) \( + k \) \( < \) \( q - 1 \) then
22. cnt(p) \( + k \) \( = \) cnt(p) \( + k \) \( + 1 \);
23. end for \( k \) from 0 to \( q - 1 \) do
24. end if
25. end for
26. return \( G' \);

Algorithm 1. ChoosePartialGrams(\( \sigma, q \))

and ii) the final candidate list must be small enough to reduce the cost of verification against the raw data, i.e., the input string \( \sigma \). These two requirements converge to index only highly selective \( q \)-grams, i.e., the ones with a small number of postings. More formally, we minimize the function \( \sum_{\gamma_i \in G} F(g_i) \), where \( F(g_i) \) denotes the frequency of \( q \)-gram \( g_i \), i.e., the number of postings in the list \( P(g_i) \). This problem can be reduced to the weighted set covering problem.

In terms of the set covering problem, the universe to be covered is all positions of the string \( \sigma \), i.e., \( p_i \in S = \{1, \ldots, n\} \), where \( n \) is the length of \( \sigma \). The collection of all posting lists \( P(g_i) \) of the \( q \)-grams \( g_i \) appearing in \( \sigma \) are used to cover the universe \( S = \{1, \ldots, n\} \). The weight of each \( P(g_i) \) is the number of elements in the list, i.e., \( F(g_i) \).

The weighted set covering problem is a well studied problem and the optimal solution can be approximated by the following greedy algorithm, within an approximation factor of \( H_n(\leq \ln n + 1) \). Let \( \gamma(g_i) = F(g_i)/|P(g_i)\cap S| \) be a weight function, which considers both the \( q \)-gram frequency \( F(g_i) \) and the number of postings \( |P(g_i)\cap S| \) to be covered next. At each step, the \( q \)-gram \( g_i \) with the minimum value of the weighted function \( \gamma(g_i) \) is picked and added to the partial index. Next, all postings that the posting list \( P(g_i) \) could remove are removed from the set \( S \). Set \( S \) contains all the uncovered positions up until this point. Finally, the function \( \gamma \) is re-computed for all remaining unused \( q \)-grams. The process is repeated by choosing the next smallest weighted \( q \)-gram, until all positions are covered, i.e., until \( S = \emptyset \).

The greedy algorithm produces a good approximation of the optimal solution in polynomial – to the size of \( \sigma \) and number of posting lists – time. However, the complexity remains quadratic, thus when both \( n \) and \( G \) are large, the classical greedy algorithm fails: it already takes hours to select 3-grams from a 64MB document.

3.1 A Scalable Set-Cover Algorithm

The complexity of the previous greedy algorithm originates from the requirement that in each step the new global minimum value of \( \gamma(g_i) \) is selected. The basic idea of Algorithm 1 is to scan all \( q \)-grams in descending frequency order and greedily choose which \( q \)-grams to keep without recomputing the global weights in each step. More specifically, a \( q \)-gram \( g_i \) is not included in the partial index if it has a large frequency \( F(g_i) \), and only if all positions can be still covered by the remaining \( q \)-grams.

Example 3.1: Going back to the example in Figure 1, the string \( \sigma = 'one\ world\ one\ dream\ one\ night\ in\ beijing' \) has 40 positions to be covered and 33 distinct 3-grams. A single pass over string \( \sigma \) (line 2) will extract all \( q \)-grams and their posting lists. The \( q \)-grams are then sorted in descending order of their frequencies (line 4). Next, a counter \( cnt(\) \) for each position in the string \( \sigma \) is initialized to 0 (line 5).

Figure 2 depicts the first few steps of the ChoosePartialGrams algorithm for the string \( \sigma \) of Figure 1. The most frequent \( q \)-gram is \( 'one' \), but it has to be selected because it covers the first position of string \( \sigma \). Thus, \( G' = \{ 'one' \} \) (lines 8-10). Next, the count \( cnt(\) \) of each position in string \( \sigma \) covered by \( 'one' \) is set to 3. The second row of Figure 2 depicts the values of the count array \( cnt(\) \) for each position in the string \( \sigma \) after 3-gram ‘one’ has been selected.

The second most frequent 3-gram is ‘\( ne_2 '\). The positions that 3-gram ‘\( ne_2 '\) covers are either already covered (e.g., position 2) or the count value \( cnt(\) \) is less than \( q - 1 \). Therefore, 3-gram ‘\( ne_2 '\’ can safely be disregarded since it is too frequent and other 3-grams do exist for covering the uncovered positions. However, before disregarding 3-gram ‘\( ne_2 '\’, each of the positions \{4, 14, 24\} is incremented by 1 to signify that there are still \( q \)-grams left that can potentially cover these positions. The third row of Figure 2 depicts the new state of the count array.

The next two 3-grams are ‘\( ne_1 '\’ and ‘\( e_1\w ' \). Both of them can be disregarded, because they are high in the frequency list and none of the positions that they cover on string \( \sigma \) have a count equal to \( q - 1 \). However, ‘\( wo\ ' \) has to be selected to ensure that position 4 of string \( \sigma \) is covered.

ChoosePartialGrams terminates with thirteen 3-grams selected for the partial 3-gram index: \( G' = \{ one, wo, rld, do, \w, dr, eam, m, o, ni, ght, \w, hi, be, yii, inq\} \).

The complexity of algorithm ChoosePartialGrams is \( O(|\sigma| + |G| |\ln|G)| \). First, a linear scan is performed on \( \sigma \) to derive all \( q \)-grams, their postings lists, and the frequencies.
3.2 Signature-based Gram Indices

There are two main limitations regarding partial q-gram indices: i) some very frequent q-grams are unavoidably selected to ensure that each byte of the string σ is covered; ii) during query evaluation, the indexed partial q-grams might not be sufficient to fully answer a query q, thus resulting in a candidate list. In order to prune the false positives from the candidate list, the string σ has to be examined, causing time consuming I/O access.

To alleviate the aforementioned limitations, we propose the use of signatures. A signature s of a q-gram g is defined to be the concatenation of the two bytes guarding g in string σ. More specifically, let a q-gram be \( g \) and its posting list be \( P(g) = \{p\} \). Then, the signature \( s \) of \( g \), in position \( p \) is \( s = \sigma[p-1]\sigma[p+q] \). A q-gram may have as many signatures as the number of postings in its posting list \( P \). The collection of all the signatures of a q-gram \( g \), together with its posting list will be referred to as an s-gram. The q-gram \( g \) is also called the infix of its s-gram. Finally, we will refer to the first byte of a signature \( s \) as \( s[1] \) and the second as \( s[2] \). If \( s[1] \) or \( s[2] \) do not correspond to a byte (i.e., they point to the positions before the start or after the end of the string \( \sigma \)) then the symbol \( \$ \) is used.

Example 3.2: Figure 3 shows the signatures \( s \) of the first two q-grams of the string \( \sigma \) of Figure 1. The 3-gram ‘one’ has the posting list \( \{1, 11, 21\} \) and the list of signatures \( \{S_{one}\} \). Consequently, the postings list of 3-gram ‘one’ can be split to two smaller ones, namely \( \{1\} \) and \( \{11, 21\} \), each one referring to a different signature of ‘one’.

Signatures are used to split long posting lists, in order to reduce both the number of postings fetched and the number of candidates produced during query processing. This is true because s-grams are more discriminative than q-grams. However, signatures may have a considerable storage overhead. In theory, the number of signatures for each q-gram is at most \( 2^{16} \) (for 2 bytes), but in practice the number of signatures for most q-grams is far below this upper bound.

The frequency distribution of q-grams in the text data sets we used loosely follows Zipf’s law. This indicates that many q-grams will have short posting lists. For these q-grams, the overhead of maintaining their signatures is higher than retrieving the whole posting list during query evaluation. Moreover, some long posting lists may be split to very fine grained pieces, as a result of a large number of different signatures. For these cases, it is preferable to combine the fine grained lists into groups and thus save storage overhead.

The next section presents techniques on how to dynamically make the decision on which posting lists to keep unmodified, which ones to split with the use of signatures and which ones to be combined into groups.

---

**Figure 3: Signature-based Posting List**

\( F(g) \) (lines 2-3). The cost of sorting all q-grams (line 4) is \( O(|G|\ln|G|) \). The big loop (lines 6-26) is a linear in the size of \( \sigma \), since each position of \( \sigma \) will be visited at most \( q \) times.

**Figure 4: An Overview of Qs-Gram**

**QS-Grams.** We propose a novel approach to marry the merits of q-gram and s-gram, named qs-gram, with the following objectives: maintain less grams on the index entries, split long posting lists for improved query performance, and merge short postings for better compression ratio.

Figure 4 gives an overview of qs-gram. Two histograms are required, \( H_s \) and \( H_q \), for s-grams and q-grams, respectively. Assume that \( q = 3 \), the number of distinct q-grams is up to \( 2^{24} \) and for s-grams \( 2^{40} \). The histogram \( H_s \) might be too large to fit in memory. A lossy technique will be discussed in Section 4 on how these histograms can be efficiently built in the secondary memory.

As depicted in Figure 4, there are less q-grams than s-grams, and in average, the frequency of a q-gram is higher than that of an s-gram. Suppose that a threshold \( t \) is used to determine whether a q-gram is frequent or not, the grams are classified as follows:

1. Frequent q-grams \( (frequency \geq t) \) will end up with a private posting list. Referring to the left part of \( H_s \) in Figure 4, they reside in an s-gram dictionary.
2. Infrequent q-grams \( (frequency < t) \) will not be split at all. These q-grams are maintained in a q-gram dictionary, as shown in Figure 4, the right part of \( H_q \).
3. For the grams in between, i.e., q-grams with frequencies no less than \( t \) and s-grams having frequencies less than \( t \), a hash-based approach is utilised.

If a q-gram \( g \) has a long posting list, but has many s-grams with short posting lists, the hash function \( h() \) is used to combine s-gram lists in a single one. To construct the hash-based q-gram dictionary, a number of buckets is dynamically allocated for each q-gram according to the frequencies of q-grams. For instance, if a q-gram \( g_q \) with frequency 35 is to be split and the threshold is \( t = 10 \), four buckets will be allocated for \( g_q \). As the hash-function randomly combines infrequent signatures, the amount of postings falling in each hash bucket approximates \( t \). The keys in the hash-based q-gram index are just q-grams stored in order, each leading to an array of buckets. To look up an infrequent s-gram, its infix q-gram \( g \) is first found, after which the hash function \( h() \) is applied on \( s \) to compute the specific bucket, as shown in Figure 4. The main reason for using a hash function to combine buckets is that it requires no additional storage space.

The qs-gram index is designed to merge short posting lists of s-grams associated with one q-gram, so as to get almost equally length \( t \) posting lists. The certainty that posting
lists have a certain minimum length ensures that even in skewed gram distributions storage volume is dominated by postings, and not by the dictionary, and also ensures that compression is functional on all postings. As twice the list length \( t \) is an upper bound on the average number of false positives, it should be chosen relatively small.

4. IMPLEMENTATION DETAILS

This section describes the implementation details for building the \( q \)-gram indices. Since the input data is larger than the available memory, we present chunk based algorithms for full and partial \( q \)-gram indices that efficiently divide the process to fit in memory. We then present how a query is evaluated against the proposed \( q \)-gram indices.

4.1 Gram Generation

We describe not only effective, but also efficient and scalable generation algorithms for full \( q \)-gram, partial \( q \)-gram and \( q\{1\} \)-gram indices. Since, the input data may not fit entirely in main memory, and in order to achieve good scalability, a chunk based generation strategy is used. Specifically, the input data is partitioned into chunks, i.e., a sequence of bytes with equal length, and each chunk is processed independently. The algorithm can be executed either sequentially or in parallel over each chunk. After the termination of the chunk based generation, the local posting lists of each chunk are merged using a multi-way merge-union operation. The size of each chunk can be tuned, in order to render the algorithm cache or memory resident.

To decrease storage overhead, as well as to improve I/O system performance, all posting lists are compressed before being written to disk. Since each posting list is sorted, delta coding [3,11] is adopted. Delta coding computes the difference between consecutive postings, and records only the inserted bytes with equal length, and each chunk is processed in-line. For full and partial \( q \)-gram and compute the frequencies, \( k \) starts at 0.

Algorithm 2 details how the local histogram and the data file is populated for each chunk. A temporary structure is used that records each \( q \)-gram and its position in the chunk \( C_i \) (line 2). First, chunk \( C_i \) is sequentially scanned and every \( q \)-gram is extracted and stored to \( \text{temp} \) (lines 3-6). Afterwards, the \( \text{temp} \) structure is sorted on the value of the \( q \)-grams by using the radix cluster algorithm [15] (a radix sort on the 8q lowest significant bits when viewing the concatenated bytes of a gram as a number) in \( O(qm) \) time (line 7). Next, the \( \text{temp} \) structure is merged to contain all the postings of the same \( q \)-gram and the frequencies are computed (line 8). This is possible, because radix cluster will group all same \( q \)-grams together. Next, the data file \( D_i \) is populated by dumping sequentially the merged and sorted \( \text{temp} \) structure (line 9). Finally, the global histogram is updated by merging the current local histogram (line 10).

The above process is repeated for all chunks. Afterwards, the local data files \( D_i \) are merged to one single global data file \( D \). Each posting list of the local data files is written to the correct offset of the single global data file by consulting the global histogram. The resulting global data file contains the entire posting list of each \( q \)-gram in the input string \( \sigma \).

An alternative approach is to scan the entire input string and collect the global histogram without writing data to disk. In a second scan, data could be written to the final global data file without merging local posting lists. However, this alternative approach writes data using random I/O, which renders the algorithm I/O bound. For the proposed approach, all disk reads/writes entail sequential I/Os, which is over an order of magnitude faster than random.

Index generation thus produces two global structures: a histogram \( H \) and a data file \( D \). The global histogram contains entries as \([q\{1\} \text{gram}]\text{data}\_\text{offset}\), where \text{data}\_\text{offset} refers to...
to the position in the data file that the \( q \)-grams’s posting list can be found. The \( q \)-grams are sorted to facilitate binary search during query evaluation. The data file is the concatenation of the posting lists of all \( q \)-grams.

The generation procedure for partial \( q \)-grams is similar to full \( q \)-grams. The only difference is that in each chunk, instead of indexing all \( q \)-grams, partial \( q \)-grams are selected using the algorithm proposed in Section 3.1. The decision is made locally by consulting only the local histograms.

For generating the \( gs \)-gram index, two global histograms are required, as depicted in Figure 4. One global histogram is needed for \( q \)-grams and one for \( s \)-grams. These are used to assist the gram classification by examining their frequencies. The two global histograms are produced in the same way as described until now. However, the \( s \)-gram histogram is an array of \(((q + 2)\)-gram, posting list\) entries sorted first on the \( s \)-gram level then on the postings. Recall that the number of \( s \)-grams is up to \( 2^{40} \) when \( q = 3 \), which could render the \( s \)-gram histogram to be out of memory. A lossy probabilistic counting technique is adopted for this case, similar to [17], where there is a limited histogram buffer. In principle all \( s \)-grams are collected, but when the buffer gets full, \( s \)-grams with a small count are removed to make space. The adaptive approach works as follows. A lower bound, initially set to 1, is used to guard whether an \( s \)-gram should be removed or not. If after removing entries with only 1 occurrence, there is still not enough space, the lower bound is doubled and the above removal operation is repeated until there is enough space in the buffer. This leads to an incomplete histogram on the \( s \)-gram level, but generally only \( s \)-grams that occur very infrequently will be missing.

After the first scan of the input string, we determine using threshold \( t \) whether a \( q \)-gram should be split with signatures: i) very frequent \( s \)-grams will end up with a private posting list, recorded in an \( s \)-gram dictionary; ii) very infrequent \( q \)-grams will not be split at all, stored in a \( q \)-gram dictionary; iii) the hash based approach is used in between.

Next, the input string is scanned for the second time. In this second pass, the local histograms and data files are generated for each chunk. We point out two differences with full \( q \)-gram generation. First, for hash based \( q \)-grams, instead of maintaining for each \( q \)-gram a global offset, each \( q \)-gram will be fetched. For the full \( q \)-gram index, the posting lists of all \( q \)-grams are indexed. However, in order to further optimize the query evaluation, \( q \)-grams with very long posting lists can be omitted during query evaluation. The basic requirement is that each byte of the query is covered at least by one \( q \)-gram. For instance, to cover a seven-byte query ‘one,world’, the \( q \)-grams ‘one’ and ‘world’ can be used. To cover the remaining byte ‘w’, any of the \( q \) possible grams (i.e., ‘ne’ , ‘e’ and ‘world’) are checked, and the one with the smallest frequency is selected.

Finally, if the \( g \)-gram is not located either in the \( s \)-gram or \( q \)-gram index, a probabilistic counting technique is adopted for this case, similar to [17], where the histograms are not sorted, therefore a binary search can be used. To cover the remaining byte ‘w’, any of the \( q \) possible grams (i.e., ‘ne’, ‘e’ and ‘world’) are checked, and the one with the smallest frequency is selected.

In the case of the partial \( q \)-gram index, unfortunately the same optimization techniques can not be applied. The \( q \)-grams are selected chunk-wise and based on local decisions (i.e., the local histogram), thus there is no guarantee that a specific \( q \)-gram will be found across the entire input string. Therefore, no matter whether a query \( q \) is fully covered or not, all posting lists must be fetched and examined.

The \( gs \)-gram index has three histograms. The infrequent \( q \)-gram histogram is processed similarly to the partial \( q \)-gram index. However, the histograms for the \( s \)-grams and hash based \( q \)-grams are treated differently, since the stored signatures must be taken into account. First, the signature \( s_i \) of each \( g \)-gram is extracted. There are three cases:

1. \( s_i = q[i - 1]q[i + q] \), where \( i \neq 1 \) and \( i \neq k - q + 1 \). The signatures are ordered, hence a binary search is used to identify the specific entry for \( s_i \).
2. \( s_i = q[k - q + 1] \), which refers to the last \( q \)-gram of query \( q \). With the use of binary search all signatures whose first byte is \( q[k - q + 1] \) will be identified.
3. \( s_i = r[q + 1] \), i.e., the first \( q \)-gram. The second byte of the signatures is not sorted, therefore a binary search can not be used and all signatures of \( g \) will be fetched.

If the \( q \)-gram is not found in the \( s \)-gram histogram, the hash based \( q \)-grams are searched. The hash value of \( s_i \) is calculated and used to identify the bucket that the \( q \)-gram should reside in. If it is found, the corresponding posting list is fetched. Notice that, if the signature contains a wildcard (for \( g_1 \) and \( g_{k-q+1} \), the hash value cannot be decided and thus all posting lists of the current \( q \)-gram will be fetched. Finally, if the \( q \)-gram is not located either in the \( s \)-gram nor the hash based histogram, the \( q \)-gram histogram is queried. All fetched relevant posting lists are then fed into a positional merge join operator in order to identify the valid
occurrences of the query $q$ in string $\sigma$. For the full $q$-gram index, a standard multi-way positional merge join is sufficient. However, for the partial $q$-gram and $qs$-gram indices, the positional merge join must adaptively determine in a chunk-wise fashion which $q$-grams are present and which were (potentially) pruned by the set-cover algorithm:

1. When processing a chunk $C_i$, if all posting lists have no postings indexed in chunk $C_i$, or some byte $g[k]$ of a query $q$ is not covered where $q - 1 < k < |\sigma| - q + 3$, chunk $C_i$ can be safely skipped.

2. If some $q$-grams have postings (are indexed) in the chunk $C_i$, and the corresponding $qs$-grams cover all bytes of the query $q$, the standard multi-way positional merge join is used.

3. Otherwise, that is, if the present $q$-grams cover the inner part of a query but not the outer boundaries, a multi-way positional merge join is still performed, but these results are marked as candidates, requiring verification afterwards.

### Extensions
Potential future optimisations in $qs$-gram query processing address the problem that when examining a chunk and seeing that a certain $q$-gram is not present, we pessimistically assumed in the above that the $q$-gram must have been pruned, while it could also be that there was simply no occurrence (in that case, there is no query result in this chunk, and whatever we emit above is a false positive). To address this, we could augment our indexing structure with a bit-string for each $q$-gram, that records whether a $q$-gram was preserved by the set-cover algorithm in a specific chunk or not. For example, the bit-string 11001 means that a $q$-gram $q$ is present in chunks 1, 2 and 5. This information can then be used to emit considerable less false positives, but raises the future research question as to the space/time trade-offs of keeping this extra information, especially on data sets with many $q$-grams. Even if no bit-strings are kept, a similar optimization can already be made in the $qs$-gram approach when merging a signature or hashed list (i.e. in case of a globally frequent $q$-gram). If this list turns out to have no postings in a chunk, we can still analyze the other hash and signature tables for the same $infix$ $q$-gram in this chunk. Because the set-covering pruning decisions are made on the $q$-gram level, presence of any other infix matching information in the indices is proof that the $q$-gram was not pruned in $C_i$, and can be used to reduce false positives. Note, however that searching for this extra proof may cause additional index I/O at query time. One potential strategy would be to use only a limited set of (or a single) long signature list(s) with the same infix to steer this optimization. Alternatively, when processing hash lists, with all hash-bucketed postings list adjacent on disk, we might simply re-use whatever information is present in the large disk block $I/O$ unit we read for opportunistic pruning.

### 5. PERFORMANCE STUDY
We conducted extensive experiments with all discussed $q$-gram indices, investigating index creation time, index space and query performance. The following datasets were used:

- wiki-article: the complete Wikipedia archive including discussion and user pages (44GB).
- XMark, using scale factors from 10 to 500 to generate XML documents from 1.1GB to 55GB.
- movie: we also experimented with a number of binary data files containing movies, and ranging in size from 1GB to 8GB, mainly to show the index creation and space behavior when there are many more different grams and the frequency is more uniform (we still lack application scenarios and queries that could be useful on such binary data, this is future work).

In the following, we use $FG$, $PG$ and $QS$ to denote full $q$-gram, partial $q$-gram and $qs$-gram indices. Most experiments were conducted on a PC with a 2.40GHz Intel Core2 Quad Q6600 CPU, 8GB of RAM and 2 hard disks in RAID-0. The experiments with the larger 426GB GOV2 dataset did not fit on the storage system of the PC, and were done instead on a server machine with two quad-core 2.8GHz Xeon (Nehalem) CPUs, 48GB of memory and a RAID-0 file system consisting of 16 SSDs (Intel X25-M). Note that in the index creation experiment, all I/O is sequential, and performance of SSDs is quite similar to normal hard drives. On this machine, we also performed a parallel experiment, where the 426GB dataset was split into eight partitions of 53GB, on which we ran our indexing program in parallel.

#### Index Creation
Table 3 shows the overall time needed to create the various indexes, where we give below each result between parenthesis the sub-component time needed for the multi-way merge union. The chunk size was 128MB for $FG$, $PG$ and $QS$ (we used frequency threshold $t = 2000$), causing a first phase of index generation for a 128MB chunk of the input data at-a-time. For the subsequent of merging local posting lists within chunk into a global index, the buffer size was again set to 128MB, causing our external merge sort to read each time 128MB of postings from all chunks, writing this out as a global postings lists. All algorithms run CPU-bound. For $FG$ and $PG$, we achieve an indexing speed of 18GB per hour on the PC, and on the server machine in the 8-way parallel experiment, the 426GB of GOV2 data was indexed in less than 3.5 hours, achieving good speedup.

For $QS$, the performance is around 8GB per hour on the PC. $SG$ and $PG$ require additional time for running the set covering algorithm, but in case of $PG$ this extra cost is offset by the fact that it writes less postings (this can be seen by comparing the write times between parentheses). $QS$ takes much more time, since (1) the radix-sort runs on $q + 2$

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Index Creation Performance (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>wiki-meta</td>
<td>wiki-article: the complete Wikipedia archive including discussion and user pages (44GB).</td>
</tr>
<tr>
<td>XMark</td>
<td>XMark, using scale factors from 10 to 500 to generate XML documents from 1.1GB to 55GB.</td>
</tr>
<tr>
<td>Movie</td>
<td>movie: we also experimented with a number of binary data files containing movies, and ranging in size from 1GB to 8GB, mainly to show the index creation and space behavior when there are many more different grams and the frequency is more uniform (we still lack application scenarios and queries that could be useful on such binary data, this is future work).</td>
</tr>
</tbody>
</table>

### Table 3: Index Creation Performance (sec)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Index Creation Performance (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>wiki-article</td>
<td>wiki-meta: the complete Wikipedia archive including discussion and user pages (44GB).</td>
</tr>
<tr>
<td>XMark</td>
<td>XMark, using scale factors from 10 to 500 to generate XML documents from 1.1GB to 55GB.</td>
</tr>
<tr>
<td>Movie</td>
<td>movie: we also experimented with a number of binary data files containing movies, and ranging in size from 1GB to 8GB, mainly to show the index creation and space behavior when there are many more different grams and the frequency is more uniform (we still lack application scenarios and queries that could be useful on such binary data, this is future work).</td>
</tr>
</tbody>
</table>

In the following, we use $FG$, $PG$ and $QS$ to denote full $q$-gram, partial $q$-gram and $qs$-gram indices. Most experiments were conducted on a PC with a 2.40GHz Intel Core2 Quad Q6600 CPU, 8GB of RAM and 2 hard disks in RAID-0. The experiments with the larger 426GB GOV2 dataset did not fit on the storage system of the PC, and were done instead on a server machine with two quad-core 2.8GHz Xeon (Nehalem) CPUs, 48GB of memory and a RAID-0 file system consisting of 16 SSDs (Intel X25-M). Note that in the index creation experiment, all I/O is sequential, and performance of SSDs is quite similar to normal hard drives. On this machine, we also performed a parallel experiment, where the 426GB dataset was split into eight partitions of 53GB, on which we ran our indexing program in parallel.
bytes instead of \( q \) bytes, (2) to save space without writing intermediate data, the raw string is scanned twice. The first scan is to collect two global dictionaries for \( s \)-gram and \( q \)-gram. The second scan processes chunk-wise data.

**Index Space.** Table 4 shows the sizes of generated indices, including the space for the dictionaries. The dictionary sizes depend on the amount of different grams (see Table 2). As the amount of different grams in the index is small in textual data, the dictionary sizes do not play any role in the results on the textual data sets in Table 4. Table 4 shows in the additional "dict," row that for \( QS \) (the scheme with the largest dictionaries), this overhead is still limited. The overall conclusion for textual data is that \( FG \) with compression typically achieves a 2x storage space, and \( PG \) and \( SG \) improve that to roughly 1x. In case of binary data, index space deteriorates to 4x in \( FG \). Both partial approaches perform similar, and reduce storage by almost a factor 2.

<table>
<thead>
<tr>
<th></th>
<th>wiki</th>
<th>wiki</th>
<th>XMark</th>
<th>GOV2</th>
<th>binary data (movie)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( FG )</td>
<td>8.3</td>
<td>11.1</td>
<td>59.6</td>
<td>39.9</td>
<td>17.6 19.1 34.0</td>
</tr>
<tr>
<td>( PG )</td>
<td>2.4</td>
<td>3.5</td>
<td>12.5</td>
<td>2.2</td>
<td>1.7  1.9  1.9</td>
</tr>
<tr>
<td>( SG )</td>
<td>0.0</td>
<td>3.0</td>
<td>0.9</td>
<td>1.0</td>
<td>0.2  2.2  3.3</td>
</tr>
<tr>
<td>dict</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0  0.0  0.0</td>
</tr>
</tbody>
</table>

Table 4: Index Storage Space (GB)

Figure 6 shows the scalability on XMark using sizes 1.1GB, up to 36GB. Figure 6 (a) confirms linear scalability in creation time and Figure 6 (b) confirms linear scalability in index sizes.

![Figure 6: XMark data with factors 10, 20 to 320](image)

**Query Evaluation** For each data set, queries with variable lengths were tested, from 5 to 15. For queries with the same length, 100 positive queries were randomly drawn from the dataset, but we excluded queries with excessively long results (> 1M). In the figures, the results are ordered on \( FG \) score to help comparison. Due to space limitations, we concentrate here on the 44GB wiki-meta data.

Figures 7 (a)-(d) shows that for short queries (e.g., 5), \( FG \) is generally faster than \( PG \) and \( QS \). However, when the length of queries increases to 9, the difference between \( FG \) and \( PG \) and \( QS \) get smaller, and \( QS \) processes many queries faster than \( FG \). For queries with lengths 11 and 15, \( QS \) outperforms \( FG \) for most queries (\( PG \) being slowest always).

To understand these results better, we first explain the other two groups of experiments. The first group is shown in Figures 7 (e)-(h), showing the number of postings loaded for different approaches. The number reported here was the posting under compression i.e. the data read from disk, but not the number of postings resulting from the positional merge join. The other group of experiments is the number of postings verified, which applies only to \( PG \) and \( QS \), since they may generate false positives. These numbers are shown in Figures 7 (i)-(l).

6. RELATED WORK

For answering exact substring matching, suffix tree \([8, 16]\) and suffix array \([9, 14]\) have been extensively studied, but this has not yet produced algorithms that could work on huge sizes such as the 426GB GOV2 dataset we indexed. \( q \)-gram based indices were first used to model sequences, using the statistical properties of \( q \)-grams. \( q \)-grams have been widely studied for efficient approximate string matching, also using DBMS as the storage and execution component \([10]\). In order to improve the performance of approximate queries, \([12, 20]\) proposed variable-length grams, and \([1, 5]\) worked on reducing the sizes of indices – note that our work concerns not approximate but with exact query processing. Many other work study the problem of approximate string joins using various similarity functions \([4, 7, 19]\).

7. CONCLUSIONS AND FUTURE WORK

We presented different gram-based indices for exact substring matching on huge data sets. Motivated by the fact that the full \( q \)-gram index has considerable storage overhead, impacting particularly the ease of manipulating such huge data sets, we aimed at space economical \( q \)-gram indices. The proposed partial \( q \)-gram indices have a very compact size (around 1x), but produce a relative large number of false positives that must be checked against the raw data. To alleviate this problem in partial \( q \)-grams, we proposed a novel approach called \( qs \)-gram index. The \( qs \)-gram index is designed to exploit the gram distribution, by splitting long posting lists and merge short posting lists with a frequency-adaptive signature approach, to ensure good data compression and query performance. The trade-off compared to partial \( q \)-grams is that \( q \)-grams consume slightly larger space. We demonstrated excellent scalability on large data sets (up to 426GB), and showed query performance of \( q \)-grams rivals or even improves that of traditional full \( q \)-grams.

As to future work, to reduce the number of postings to
be verified in query processing, we plan to build a cost model to judge when is benefit to load and combine multiple posting lists instead of verifying the raw data. Also, the dictionary size (and auxiliary bit-strings) for large binary data is problematic. We will investigate dictionary compression techniques.

8. REFERENCES


