Zwift: A Programming Framework for High Performance Text Analytics on Compressed Data

Feng Zhang [†‡⋄], Jidong Zhai [⋄], Xipeng Shen [#], Onur Mutlu [⋆], Wenguang Chen [⋄]
[†]Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China), MOE, China
[‡]School of Information, Renmin University of China, China
[⋄]Department of Computer Science and Technology, Tsinghua University, China
[#]Computer Science Department, North Carolina State University, USA
[⋆]Department of Computer Science, ETH Zurich, Switzerland
fengzhang@ruc.edu.cn, zhaijidong@tsinghua.edu.cn, xshen5@ncsu.edu, onur.mutlu@inf.ethz.ch, cwg@tsinghua.edu.cn

ABSTRACT

Today's rapidly growing document volumes pose pressing challenges to modern document analytics frameworks, in both space usage and processing time. Recently a promising method, namely *text analytics directly on compressed data (TADOC)*, was proposed for improving both the time and space efficiency of text analytics. The main idea of the technique is to enable direct document analytics on compressed data. This paper focuses on the programming challenges for developing efficient TADOC programs. It presents Zwift, the first programming framework for TADOC, which consists of a Domain Specific Language, a compiler and runtime, and a utility library. Experiments show that Zwift significantly improves programming productivity, while effectively unleashing the power of TADOC, producing code that reduces storage usage by 90.8% and execution time by 41.0% on six text analytics problems.

CCS CONCEPTS

• **Software and its engineering** → **Compilers**; **Domain specific languages**; • **Information systems** → *Data analytics*;

KEYWORDS

Compiler, Domain Specific Language, Text Analytics

1 INTRODUCTION

As data analytics becomes an increasingly important workload running on today's HPC systems, how to improve the performance of high-performance data analytics is already an important topic in HPC [10, 20, 31]. Text analytics is an important class of high performance data analytics problems, specializing on deriving statistics, insights, or knowledge from textual documents, such as system log files, textual content of web pages, phone or email records. It is essential for many domains, ranging from HPC system diagnosis, to health, security, and more.

This paper explores a recently proposed approach to improving both the time and space efficiency of text analytics, namely *text analytics directly on compressed data (TADOC)* [30]. The basic idea of TADOC is to conduct text analytics on compressed data directly, without recovering the original uncompressed data.

Even though data compression is commonly used in text analytics, all existing practices need to first decompress the data before processing them, which does not reduce but lengthen the processing time. An exception is Succinct [2]. Although Succinct processes compressed data directly, it is specific to the database domain, designed only for searching and random accessing for a specific query, rather than supporting general text analytics (more in Section 8).

By leveraging a compression algorithm called Sequitur [19], TADOC is feasible for general text analytics. Sequitur compresses a sequence of symbols into a Context-Free Grammar (CFG), which can be represented with a Directed Acyclic Graph (DAG). Text analytics can then be performed directly on the DAG. Compression by Sequitur takes time. Fortunately, many datasets (e.g., Wikipedia [1], UCI Machine Learning Repository [15]) are used for various analytics tasks by many users repeatedly. For them, compression time is well justified by the repeatedly usage of the compression results.

Directly working on compressed data, TADOC may save both time and space. Space-wise, its reduced data size saves not only the storage space, but also the amount of memory needed for processing the data. Time-wise, because it processes the compressed data where multiple duplications of an entry in the original dataset are already folded into one copy, TADOC can possibly avoid repeatedly processing those duplications, and hence achieve speedup.

However, developing programs to effectively apply TADOC to a given text analytics problem is challenging, especially for general programmers. The key for TADOC's time benefits is its reuse of processing results across multiple appearances of the same string in the original input. That helps avoid repeated processing. However, such reuse requires the computation results to be saved and frequently propagated through the DAG. It is hence both important and tricky to strike a good tradeoff between the reuse and the overhead.

The problem is further complicated by the differences between the various analytics problems (e.g., some care about the boundaries of different files in the input dataset, some care about the order of words), the attributes of input datasets (e.g., sizes, numbers of files, unique words), and the demands for scalable performance. As a result, suitable ways (data structures, DAG traversal order, processing algorithms) to implement TADOC differ across analytics problems

ICS '18, June 12–15, 2018, Beijing, China

ACM ISBN 978-1-4503-5783-8/18/06…$15.00
https://doi.org/10.1145/3205289.3205325

as well as across input datasets. Although TADOC converts document analytics to graph problems, using general graph processing frameworks (e.g., Ligra [25]) does not solve these problems as these complexities are specific to text analytics; programmers would still need to deal with these complexities in their own code. A previous study [30] discusses the complexities and provides some guidelines for programmers to follow. However, developing efficient TADOC remains a difficult task for general programmers.

The objective of this work is to remove the major barriers for practical adoption of TADOC through the development of a programming system support. Specifically, we present Zwift, the first programming framework for TADOC. Zwift consists of an embedded domain-specific language (DSL), called Zwift Language, a compiler and runtime, and a utility library.

Using the Zwift Language, programmers can easily express the basic elements of the text analytics of interest. The language also offers programmers the flexibility to write multiple (optional) ways to process the data if they are uncertain about which one works best. With the Zwift Language, programmers can concentrate on the functionality aspect of the data analytics solution without having to consider the many performance factors.

The Zwift compiler automatically assembles these basic elements into an efficient program for TADOC. A major challenge is how to make the produced code able to employ the data structures and algorithms that best fit the analytics problems. Zwift addresses this challenge via a hybrid method that combines rules with offline profiling. Users have the option to provide multiple versions of code and some training inputs, and Zwift compiler tries to generate related versions, measures the time on the training input, and selects the suitable version.

Zwift relieves programmers from most of the efficiency concerns and complexities for implementing TADOC. Our experiments on six text analytics problems demonstrate that Zwift significantly improves programming productivity; it reduces the the number of lines of source code by 84.3%, and the program development time by 80%. At the same time, Zwift-produced code has performance that is comparable to or even better than that of manually-developed code. Overall, Zwift is effective in unleashing the power of TADOC: it reduces storage usage by 90.8% and execution time by 41.0%, compared to data analytics on uncompressed data.

2 BACKGROUND

This section presents the background of TADOC, including the compression algorithm it builds on, and its basic idea.

2.1 Sequitur

The underlying compression algorithm of TADOC is Sequitur. Sequitur [19] is an algorithm proposed by Craig Nevill-Manning and Ian H. Witten in 1997 for inferring a hierarchical structure from a sequence of discrete symbols. For a given sequence of symbols, it derives a context-free grammar (CFG), with each rule in the CFG reducing a repeatedly appearing string into a single rule ID. As it references all occurrences of the original string with the corresponding rule ID, the resulting CFG is usually more compact than the original input.

Figure 1 gives an illustration. Figure 1(a) shows the original input, and Figure 1(b) shows the output of Sequitur in form of a CFG. The CFG uncovers the repetitions in the input string as well as the hierarchical structure of the string. It uses R0 to represent the entire string, which consists of substrings represented by R1 and R2. The two instances of R2 in R0 reflect the repetition of "a b c a b d" in the input string. Similarly, the two instances of R1 in R2 reflect the repetition of "a b" in the substring of R2. The output of Sequitur is often visualized with a directed acyclic graph (DAG) as Figure 1(c) shows. The edges indicate the hierarchical relations among the rules. For efficiency, in Sequitur compression results, we represent each word with a unique non-negative integer, and each rule ID with a unique integer greater than N, where N is the total number of unique words contained in the dataset. Figure 1(d) gives the numeric representations of the words and rules in Figures 1(a,b), while Figure 1(e) shows the CFG in the numerical IDs.

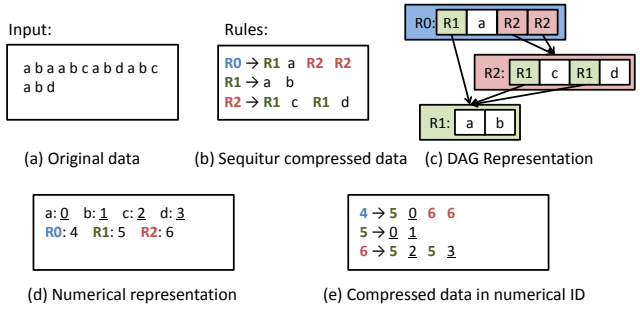

Figure 1: Illustration of Sequitur for compression.

Sequitur has several properties that make it appealing for our use. First, the CFG structure in its results makes it easy to find repetitions in input strings. Second, its output consists of the direct (sequences of) input symbols rather than other indirect coding of the input (e.g., the distance used in LZ77 [33] and suffix array [18]). These properties make Sequitur a relatively easy fit for materializing the idea of compression-based document analytics. With these said, we do not rule out the possible feasibility on other compression algorithms. The insights attained in this work hopefully could give some enlightenments when other compression algorithms are explored in the future. A limitation of Sequitur is that its compression is relatively slow. However, as TADOC is designed for data with repeated usage, compression time is not a main concern.

2.2 Basic Idea of TADOC

We take *word count* as an example to explain the basic idea of TADOC.

Word count [3, 6, 21] is a basic algorithm widely used in document classification, clustering, and theme identification. It counts the total appearances of every word in a given dataset which often consists of a number of files.

- Input: {file1, file2, file3, file4, file5, ...}
- Output: <word1, count1>, <word2, count2>, ...

Figure 2 shows how TADOC may implement *word count*. It uses the DAG that Sequitur generates on the string shown in Figure 1 (a) as the example. The processing consists of two main steps. The

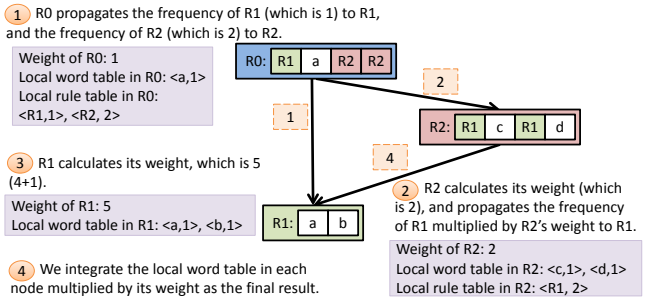

Figure 2: Illustration of a preorder traversal algorithm for `word count` on the same string as Figure 1 deals with. The weights on an edge between two nodes (say, $node_x$ to $node_y$) indicate the total appearances of $node_y$'s rule incurred due to all appearances of $node_x$'s rule.

first step calculates the total frequencies that each rule appears in the entire dataset (① to ③ in Figure 2). This step is done through a preorder traversal of the graph: Parent nodes pass their total frequencies to their children, from which, the child nodes can calculate their total frequencies. With this step done, the second step (④ in Figure 2) just needs to go through each of the rules once (with no need for another graph traversal). When it goes through a rule, it can use the total frequencies of the rule to calculate the local frequencies of a word. Here, the local frequency of a word is the directly observable frequency of the word in the right-hand side of the rule (i.e., without considering its subrules), multiplied with the frequency of that rule. By summing up the local frequencies across all rules for each word, this step gets the final answer.

Note that there are other possible implementations. For example, we can traverse the graph in a postorder (children before parents): Each node counts local word frequency (i.e., the word counts immediately contained in this node's rule without considering subrules) and then adds the word counts from its children nodes to its own word counts. The word counts obtained in the root node give the final results.

3 PROGRAMMING CHALLENGES

Developing an efficient program for effectively applying TADOC to a given text analytic problem is challenging, especially for general programmers. It fundamentally faces a cost-benefit tradeoff that is determined by many factors, ranging from what data structures to use for saving and propagating information across the DAG, to the traversal order of the DAG, the different properties and requirements of different text analytics problems, the attributes of the input datasets, and so on. The seminal study of TADOC [30] describes complexities in three main dimensions, which we list as follows.

(1) *Problem dimension.* Different analytics problems have different requirements and complexities. We use two other analytics problems to explain.

Inverted index [3, 8, 27] builds word-to-file indices for a document dataset. It is widely used in search engines. The input of *inverted index* is a set of files, and the output is the mappings between words and files. So unlike *word count* which treats inputs as a bag of words, *inverted index* distinguishes one file from another.

Sequence count [3, 14, 29] is another example problem. It counts the number of appearances of every l-word (e.g., l=3) sequence in each file. *Sequence count* is very useful in semantic, expression, and sequence analysis. Compared to *word count* and *inverted index*, *sequence count* needs to not only distinguish between different files, but also discern the orders of consecutive words.

Due to these special requirements, the algorithm in Figure 1 cannot work for these two problems. The programmer must redesign the algorithm to somehow identify and carefully handle the file boundaries. For *sequence count*, the algorithm must, in addition, beware the appearing order of words in the DAG. As an l-word sequence may span the boundaries of multiple nodes in the DAG, the algorithm must treat the boundaries and order of nodes with caution.

(2) *Implementation dimension.* As shown in the *word count* example in Section 2.2, there are multiple ways to implement TADOC for a given data analytic problem; they may use different order for graph traversal, and employ different data structures to store and pass intermediate data across the DAG. As a result, they may save different amounts of processing time and also incur different amounts of overhead. The best decision may be hard to determine as it may depend on the attributes of the input datasets (as discussed next). In addition, there are different types of execution platforms, ranging from parallel to distributed environments. Writing parallel or distributed code is often difficult for general programmers; requiring a version for each type of platform adds more difficulty for the development and maintenance.

(3) *Dataset dimension.* The attributes of input datasets may sometimes substantially affect the overhead and benefits of a TADOC implementation. For instance, when solving *inverted index*, one method is to propagate through the graph the list of files in which a word appears. It could work if there are a modest number of files, but would incur substantial propagation overhead when the number of files is big as the list to propagate could get very large. So datasets of different properties could demand a different design in what to propagate and the overall traversal algorithm.

These complexities make it challenging for general programmers to adopt TADOC. Our experience shows that a poorly written TADOC could easily make the overhead outweigh the benefits. These difficulties motivate the development of Zwift.

4 A CONCEPTUAL FRAMEWORK

To create a programming framework for TADOC, we need to first get a unified view at the various TADOC problems. It could allow us to map the many different problems to the same abstraction, and hence offering a unified programming framework.

The view we get is inspired by the classic *data flow framework* in compilers which unifies the treatment to various data flow problems. Through examinations of various text analytics problems, we create *a unified conceptual framework of TADOC* (*unified framework* in short). It captures the common operations and workflows of TADOC through a six-element tuple $(G, V, F, \bigvee, D, \bigwedge)$, defined as follows:

(1) A graph G, which is the DAG representation of the compression results of Sequitur on a dataset

(2) A domain of values V, which defines the possible values as the outputs of the document analytics of interest
(3) A domain of values F, which defines the possible values to be propagated across G
(4) A meet operator $\bigvee$, which defines transformations of values in F at one step of the traversal of G
(5) A direction of the value flow D, which is either FORWARDS (parents before children) or BACKWARDS (children after parents) or DEPTHFIRST (in the order of the original files)
(6) A gleaning operator $\bigwedge$, which defines transformations of values in V in the final stage of the analytics.

The solutions to the various document analytics problems listed in the previous sections can all be mapped to this abstract model. Each can be regarded as an instance of the *unified framework*, with the six components instantiated with the values specific to the analytics problem.

Take `word count` as an example. For that problem, G is the DAG of the Sequitur compression results on the dataset of interest, and V is the set of $< w, n >$, where w is a possible word, and n is a non-negative integer. The two algorithms described in Section 2.2 for solving the `word count` problem are essentially different instances of the *unified framework* with different definitions of F, D, $\bigvee$, and $\bigwedge$. The solutions to the other two analytics problems outlined in Section 3 can be mapped to the *unified framework* in a similar manner. For sequence counting, for instance, G is the DAG of Sequitur results, V is $< s, n >$ (s for l-long word sequences, n a non-negative integer), F contains empty operation, D can be DEPTHFIRST, $\bigvee$ increases counts in some global or local tables to record intermediate results, $\bigwedge$ calculates frequencies of rules and combine recorded results when necessary.

High-Level Algorithms for Solving Sequitur-Based Document Analytics Problems. After abstracting various concrete analytics problems into our *unified framework*, we can examine the solutions to various Sequitur-based document analytics problems in a single view. Despite their differences, all the solutions to these problems are essentially variants of the following high-level algorithm.

`High-level algorithm for Sequitur-based document analytics:`

(1) *Loading:* Load Sequitur grammar, build G (or its variants), and initiate the data structures local to each node or global to the algorithm.
(2) *Propagation:* Propagate information with the meet operator $\bigvee$ while traversing G in direction D.
(3) *Gleaning:* Glean the information through the gleaning operator $\bigwedge$ and output the final results.

We found that this high-level algorithm suited all document analytics problems we have encountered. The main differences among the different solutions to a problem are mostly about the information they propagate, the meet and glean operators they use, the direction of the traversal, and the assistant data structures they employ. These aspects together cover the set of most important factors that determine the efficiency of the constructed solutions, and lay the basis for our design of the Zwift programming framework.

5 LANGUAGE ZWIFT

Based on the conceptual framework in the previous section, we develop a DSL named Zwift Language for programming text analytics to make use of TADOC with ease. To ease understanding by the HPC/Analytics community, we try to explain the DSL through informal intuitive descriptions and examples rather than formal language semantic definitions. Listing 1 shows the main constructs in Zwift.

Listing 1: Zwift DSL template.

```
1  ELEMENT = LETTER/WORD/SENTENCE
2  USING_FILE = true/false
3  NodeStructure = {
4  //data structures in each node
5  }
6  Init = {
7  //initializations for each node in ZwiftDAG
8  }
9  Direction = FORWARDS/BACKWARDS/DEPTHFIRST
10 Action = {
11 //actions taken at each node during a traversal
12 }
13 Result = {
14 // result structure
15 }
16 FinalStage = {
17 // final operations to produce the results
18 }
```

These constructs correspond to the elements in the *unified framework* presented in the previous section.

(1) `ZwiftDAG` represents the DAG from Sequitur, corresponding to G in the *unified framework*.
(2) `Result` represents the structure of the analytics result, which corresponds to V in the *unified framework*.
(3) `NodeStructure` represents the user-defined data structures in each node (F in the *unified framework*). The construct `Init` is for programmers to define the initializations of the data structures in each node.
(4) `Action` represents the operations to take in each node during the traversal of the DAG ($\bigvee$ in the *unified framework*).
(5) `Direction` represents the direction for data propagation (D in the *unified framework*).
(6) `FinalStage` defines how to get the final result (the gleaning operator $\bigwedge$ in the *unified framework*). Zwift has an optional section, FinalMerging, which is about how to merge the results from FinalStage when the inputs are partitioned into multiple chunks.

Currently, Zwift is implemented as a DSL embedded in C/C++. It allows only C/C++ code inside the Zwift code constructs (e.g., `Init`, `Action`, `FinalStage`). Adding support of Python or other programming languages is future work.

In addition, Zwift has some reserved keywords, shown in Table 1. These keywords are in four categories. The first category is for indicators. Users set these keywords to indicate certain attributes of the analytics problem (e.g., whether the analytics discern file boundaries). The second category is data types. Some of them are abstract data structures, which get instantiated by the compiler and runtime based on the attributes of the analytics problem and the uncovered properties of the input datasets. An example is `ZwiftVec`, which could be substantiated with a normal C++ vector, a one-level bit vector, or a two-level bit vector as Section 6.2 elaborates. The third category is for the keywords for replacement. Zwift compiler

Table 1: Some Keywords in Zwift.

Type	KeyWord	Description
indicators	ELEMENT	granularity of input: LETTER/WORD/SETENCE, it is also a data type
	USING_FILE	whether the DAG needs to include file information; default: false
	ZwiftNodesExceptRoot	if true, the root node shall not perform the codes in `Action`; default: false
	NODE_COARSENING	node coarsening threshold; default: 0 for no coarsening
	EDGE_COARSENING	whether Zwift performs edge coarsening; default: false
	INPUT_PATH	the directory path of the compressed data; default: the first command-line parameter
data type	RULEID	data type of rule ID
	FILEID	data type of file ID
	ZwiftMap	Zwift map data structure
	ZwiftVec	Zwift vector data structure
	ZwiftSet	Zwift set data structure
replacement	ROOT	the root rule ID
	NODE	the node itself, this will be replaced by related codes in compiler
	CHILD	child node, this key word will be replaced in compiler
	PARENT	parent node, this will be replaced in compiler
built-in variables	ZwiftDAG	the DAG representation of the compression results
	FILES	number of files
	RULES	number of rules
	WORDS	number of words
	ruleID	the ID of the rule, use NODE.ruleID to get it

replaces these keywords in compilation. The fourth category consists of some built-in variables, such as the number of files and so on. Users code can use these structures without defining them; the compiler inserts code to define their values at runtime.

We take `word count` as an example to show the Zwift code in Listing 2. It first indicates that the application works at word level and does not need to discern file boundaries. It then gives the compiler the permission to optimize the DAG by coarsening its edges (explained later in this section). It defines `NodeStructure` as a structure consisting of a `wordCount` table, a `ruleCount` table, and an integer field named "weight". Note that in compilation, the Zwift compiler automatically adds some extra fields into `NodeStructure`, including an integer field `ruleID` corresponding to the rule ID of the node, and several other fields for assisting the runtime traversal order of the nodes in the DAG, as the next section will explain. `NodeStructures` are initialized through the `Init` block, where, the keyword `NODE` indicates the node itself and will be replaced by the compiler with actual variables, such as "ZwiftDAG.Nodes[i_zwift]". `Zwift_IsWord()` checks whether an element in a rule is a word or a rule. The code sets the DAG transversal direction to `FORWARDS`. Zwift has three predefined traversal directions; `FORWARDS` means a preorder (parents before children), `BACKWARDS` means a postorder (children before parents), `DEPTHFIRST` means a depth-first traversal of the DAG. The `Action` block specifies the actions for each node to take during the DAG traversal. The construct `Result` defines the structure of the final output. The construct `FinalStage` specifies the actions to take at the end of the DAG traversal.

Zwift has an optional section of `GlobalVariables` for programmers to define some global variables, and a related section of `GlobalVariableInit` for initializing those variables.

When the programmer does not know the optimal implementation (the optimal may be input dependent), she can provide several versions of Zwift code for constructs `Init`, `Direction`, `Action`, and `FinalStage`. The version numbers are expressed as subscripts of the constructs (e.g., `Action[0]` and `Action[1]`). Next section will explain how Zwift framework selects the appropriate versions on the fly.

Listing 2: Word Count Example using Zwift

```
1  ELEMENT = WORD
2  USING_FILE = false
3  EDGE_COARSENING = true
4  NodeStructure = {
5    ZwiftMap<ELEMENT, int> wordCount;
6    ZwiftMap<RULEID, int> ruleCount;
7    int weight;
8  }
9  Init = {
10   if (NODE.ruleID == ROOT) {
11     NODE.weight = 1;
12   }else{
13     NODE.weight = 0
14   }
15   for(vector<int>::iterator i=NODE.input.begin();
16       i!=NODE.input.end();
17       i++) {
18     if( Zwift_IsWord(*i) )
19       NODE.wordCount[*i]++;
20     else
21       NODE.ruleCount[*i]++;
22   }
23 }
24 Direction = FORWARDS
25 Action = {
26   CHILD.weight +=
27     PARENT.weight * PARENT.ruleCount[CHILD.ruleID];
28 }
29 Result = {
30   ZwiftMap<ELEMENT, int> result;
31 }
32 FinalStage = {
33   for(int i=0; i<RULES; i++){
34     for(ZwiftMap<ELEMENT, int>::iterator
35         j = ZwiftDAG.Nodes[i].wordCount.begin();
36         j != ZwiftDAG.Nodes[i].wordCount.end();
37         j++)
38       result[j->first]+=(j->second)*ZwiftDAG.Nodes[i].weight;
39   }
40 }
```

6 ZWIFT COMPILER AND RUNTIME

The Zwift compiler and runtime play a critical role in creating efficient sequential, parallel, and distributed code for applying TADOC based on the Zwift code. This section first gives an overview of the compiler and explains its basic functionality of assembling the Zwift code together. It then elaborates on the important optimizations conducted by the compiler and runtime through data abstraction, coarse-grained partition, and version selection.

6.1 Overall Structure and Basic Functionality

The Zwift compiler takes in the code in the Zwift Language and generates the complete C/C++ code. It is based on code skeletons we have constructed based on the common patterns of TADOC programs. Listing 3 outlines the excerpted code skeleton for the sequential version. There are skeletons for parallel and distributed code in Pthreads and Spark respectively. At a high level, the compilation goes through the following four main steps: (1) The compiler reads in the Zwift code and performs static analysis, and replaces some of the keywords (e.g., ROOT, NODE), instantiates the abstract data structures (e.g., ZwiftVec, ZwiftMap), and completes the data structures and code blocks in the Zwift code (e.g., adding some counters into NodeStructure for dependence tracking as explained later); (2) The compiler inserts the completed code blocks into the code skeleton that conforms the traversal direction specified in the Zwift code; (3) The compiler converts the code to parallel and distributed versions by leveraging their corresponding code skeletons. When there are multiple choices of the `Action` code blocks, the compiler generates multiple versions of the program with each version corresponding to one of the choices. (4) The programmers can do profiling runs on the multiple versions and create a version selector for runtime adaptations.

Some of the operations are worth further explanations, in particularly, the treatments to different traversal orders in the code generation process, the instantiation of the abstract data structures, the version selection, and the creation of the parallel and distributed versions. We explain the traversal order part first and then use separate subsections to elaborate on the other aspects.

Zwift currently has three predefined traversal directions of DAG, forward, backward, and depth-first. The compiler generates respective code to materialize each of them.

Forward traversal requires a preorder (parents before children). The Zwift compiler inserts extra fields into `NodeStructure` and employs a ready queue to efficiently ensure the correct visiting order of nodes. Specifically, it adds a field `INEDGES` and a field `CNT` into `NodeStructure`. At graph loading time, the `INEDGES` of each node is set to the total number of incoming edges (i.e., parents) of the node. During the graph traversal, updates to the `CNT` field of a node track the number of its parents that have been processed. Once it equals `INEDGES`, this node is put into the ready queue. Only nodes in the ready queue can get processed. Initially, only the root node is in the queue. A node is removed from the queue after it is processed. When the queue is empty, the traversal is done.

In Zwift DAG, a parent node knows which nodes are its children but a child node has no knowledge about its parents. The backward traversal is hence implemented in recursive calls. The calls start from the root node and recursively reach the leaves where the actual processing gets started. Because a node can be the child of multiple parent nodes, Zwift compiler inserts code to track the status of a node at runtime to avoid processing a node repeatedly.

Listing 3: A code skeleton used by Zwift compiler.

```
#include"Zwift.h"
...
struct NodeStructure{
//  <- insert the NodeStructure in Zwift code here
  ...
}
void init(){
  ...
  for(vector<int>::iterator i_zwift = ZwiftNodes.begin();
    i_zwift != ZwiftNodes.end();
    i_zwift++) {
//  <- insert the init in the Zwift code here
  }
  ...
}
// Zwift provides different functions for action,
// which relates to directions, such as
// ForwardAction(), BackwardAction(). We take
// ForwardAction() as an example
void ForwardAction(){
  ...
// construct for-loop, using Listing 2 as example
  for(ZwiftMap<int, int>:: iterator i_zwift
      = ZwiftNodes[head].ruleCount.begin();
    i_zwift=ZwiftNodes[head].ruleCount.end();
    i_zwift++){
//  <- insert the action in the Zwift code here
  ...
  }
  ...
}
void FinalStage(){
//  <- insert the FinalStage in Zwift code here
}
int main() {
  ...
  init();
  ForwardAction();
  FinalStage();
  ...
}
```

The depth-first traversal is particularly useful for analytics (e.g., `sequence count`) that are sensitive to the original appearing order of words in the input. It ensures that the traversal order is in the same order of the appearances of words in the original input to facilitate the treatment to sequences spanning node boundaries. The depth-first traversal is implemented by recursively processing each item in each rule. Reuse intermediate results across nodes is still possible. For example, in `sequence count`, the counting results of the word sequences within a node are stored in a local data structure for later reuse.

As Zwift takes care of all these implementation details, it allows programmers to focus only on the primary functionality aspects of the analytics problem. We next describe some key optimizations by the Zwift compiler and runtime for a high computing efficiency.

6.2 Data Abstractions

One feature of Zwift is its leverage of abstract data structures for compile-time and run-time optimizations. Programmers use these abstract data structures in their Zwift code without worrying about their efficiency. The compiler instantiates them with one of several optional concrete data structures during code transformations. It makes the selection based on the attributes of the analytics problem.

It meanwhile inserts code such that at runtime, the code automatically configures the use of these concrete data structures based on the properties of the input datasets to maximize the computing performance. We explain the optimization through two of the abstract data structures, `ZwiftVec` and `ZwiftDAG`.

ZwiftVec. A `ZwiftVec` can be instantiated to a normal vector, a one-level bit vector, or a two-level bit vector. The compiler makes the decision based on the problem attributes specified in the given Zwift code for high efficiency. It uses the bit vectors when a large set is needed to record boolean information. For example, for file-sensitive problems such as `inverted indexing`, the algorithm needs to propagate across the DAG the set of files that contain each word. One possible design is that each node uses a *set* structure to store the IDs of the files containing a certain word. Frequent queries and insertions to the set could cause large overhead to the processing. An alternative design is that each node uses a bit vector, with each bit corresponding to a file: 1 for file presence, 0 otherwise. It can help replace the slow set operations with fast bit operations. However, it does not work for all cases. When the number of files is huge, this design could incur large space overhead, as every node needs such a vector, and the length of each of the vectors needs to be the same as the total number of files.

Zwift addresses the problem by adopting a double-layered bitmap, illustrated in Figure 3. Level two contains a number of N-bit vectors (where N is a configurable parameter). Level one contains a pointer array and a level-1 bit vector. The pointer array stores the starting address of the level-2 bit vectors, while the level-1 bit vector is used for fast checks to determine which bit vector in level 2 contains the bit corresponding to the file in querying. If a rule is contained in only a subset of files, whose bits correspond to some bits in several level-2 vectors, then the level two of the bitmap associated with that rule would contain only those several vectors, and most elements in its pointer array would be null. The two-level bit vector adds more indirect references than a one-level vector does. But it uses much less space. The number of elements in the first level arrays and vectors of the double-layered map is only $1/N$ of the number of files.

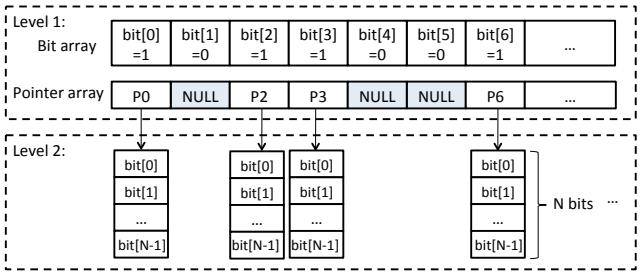

Figure 3: An illustration of double-layered bitmap for both memory footprint minimization and access efficiency.

Zwift includes all the three kinds of implementations for `ZwiftVec`. At compile time, it decides whether a normal vector is going to be used based on whether the vector is used as a set for boolean information. If that is the case, it postpones the decision of using one level or two-level bit vectors to the Zwift runtime. It does that by creating multiple versions of the code with each using a different bit vector implementations for a `ZwiftVec`. At runtime, after the graph is loaded, the runtime calculates the space cost of using one-level bit vector based on the graph size, and if the cost is over a threshold (e.g., half of the available memory), it selects the two-level bit vector version.

ZwiftDAG. `ZwiftDAG`, the abstract data structure for the loaded DAG, features an optimization, *coarsening*, which includes edge merging and node coarsening.

As an abstract data structure that programmers use in their Zwift code, `ZwiftDAG` is a data structure to hold the loaded DAG. It gets materialized at compile time. The structure of its node is determined by `NodeStructure` in Zwift code (after the compiler adds extra fields). In user's code, users can use `ZwiftDAG.Nodes` and `ZwiftDAG.Edges` to represent all the nodes and edges in the DAG. For example, Listing 2 uses `ZwiftDAG.Nodes` in summarizing the word counts in `FinalStage`.

Coarsening is an optimization happening during the graph loading time. Through it, the nodes or edges in the Zwift can represent the accumulated information of a set of nodes or edges. The two coarsening methods, edge merging and node coarsening, can be used together by Zwift runtime. `Edge merging` merges multiple edges between two nodes into one, and a weight of the edge may be used to indicate the number of original edges in the Zwift code. Edge merging loses the order of words, but helps reduce the size of the graph and hence the number of memory accesses in the graph traversal. It is helpful to analytics tasks that are insensitive to word order (e.g., `word count` and `inverted index` but not `sequence count`). `Node coarsening` inlines the content of some small rules (which represent short strings) into their parent rules; those small nodes can then be removed from the graph. It reduces the size of the graph, and at the same time, reduces the number of substrings spanning across nodes, which is a benefit especially important for analytics on word sequences (e.g., *sequence count*).

Coarsening adds some extra operations, but the time overhead is negligible if it is performed during the loading process of the DAG. On the other hand, it can save memory usage and graph traversal time.

The Zwift Language provides constructs, `NODE_COARSENING` and `EDGE_COARSENING`, to allow users to indicate their preferences for coarsening as shown in Table 1. Our compiler automatically generates the code to load users' data files to instantiate ZwiftDAG, and does the coarsening at loading time based on users' specifications. To enable edge coarsening, user sets `EDGE_COARSENING` to `true`; to enable node coarsening, users need to set `NODE_COARSENING` to a positive number, which serves as the threshold of the minimum size of a node in terms of the length of its contained string. For nodes below that threshold, their contents are folded into their parent nodes at the graph loading process, starting from leaf nodes.

6.3 Parallel and Distributed Code Generation

To get scalable performance, Zwift generates parallel and distributed code versions to leverage parallel and distributed computing resources. We have explored two methods for parallelism. The first method is a fine-grained method to distribute rules to different threads, which mainly parallelize the `Action` part in the Zwift code.

For instance, in the forward processing, the compiler could generate code to allow multiple threads to process the nodes in the ready queue simultaneously. This method did not give us significant parallel performance, due to the overhead in synchronizations and the limited amount of parallelism. Some sophisticated parallel graph processing techniques [30] may potentially help; in-depth explorations are left for future work.

The second method explores coarse-grained parallelism, which shows much better results. The parallelism is at the data level. When users use the Zwift utility library (described in Section 6.5) to compress data, by setting some parameters of the call, they can let the compression API to first partition the input data into some partitions of similar sizes, which can then be processed by different threads or processes in parallel. To take advantage of this feature, programmers need to add one extra construct, `FinalMerging`, into the Zwift program besides the standard parts . This construct specifies how to merge the processing results of the partitions into the final output.

The coarse-grained method may lose some opportunities for reusing the results as the reuse happens only within each partition. However, it avoids the frequent synchronizations the fine-grained method suffers from. Because typical text analytics use large datasets, our experiments show that the coarse-grained method usually gives better and more scalable performance. On `word count`, for instance, the fine-grained parallel version runs 50% slower than the coarse-grained version even on a small dataset (dataset A in Table 4).

This method also simplifies the conversion from the sequential code to parallel and distributed versions of the code. Because there are no communications needed between the processing of different partitions, the compiler just needs to let each thread or process run the same sequential code on one partition, and calls the merging block to conduct the final merging. Our implementation uses a skeleton in Pthreads for the parallel version and a skeleton in Spark for the distributed version. For the Spark version, Zwift uses Spark pipe() [28] to directly call the sequential version.

We note that a side benefit of the coarse-grained method is that it also allows parallel compressions of the partitions of the input data, which helps shorten the compression process.

6.4 Version Selection

As explained in Section 5, for a given application, programmers may provide multiple Zwift `ACTION` code versions with different DAG traversal orders and implementations for an analytic problem. Based on the source code, it can be hard to choose the best implementation, because the performance of some applications depends on the input datasets.

The solution currently employed in Zwift is a hybrid method that combines rules with offline profiling. The primary rule is that if the problem is an order-sensitive problem, then `DEPTHFIRST` should be the only choice. This rule comes from the earlier discussions on the special needs of such problems. Other cases are addressed currently through a synergy between Zwift and programmers. By offering the support for automatically generating multiple versions of the program (as aforementioned), Zwift allows programmers to easily run the different versions on some representative inputs. After observing the performance of those versions, the programmer may either settle on one of the versions or write a wrapper to check certain conditions of the given input dataset to select the appropriate version to run.

A potential improvement is to automate the version selections by deriving some analytical performance models of the various versions of `ACTION` on DAGs, and then, at runtime determine the best version to use by plugging into the performance models the actual features of the current input datasets. We leave a thorough exploration of this direction to the future.

6.5 Extra Optimization

An extra optimization we find helpful is to do a further compression (e.g., Gzip) upon the compressed results from Sequitur. That adds some time overhead for data processing as the Sequitur results need to be first restored before TADOC starts. However, we find that the time overhead is orders of magnitude smaller than restoring the original data because Sequitur results are much more compact than the original data. Meanwhile, this optimization saves space significantly as Section 7 shows.

7 EVALUATION

In this Section, we use six benchmarks and five datasets to evaluate the performance of TADOC and the effectiveness of Zwift in supporting TADOC, in terms of both programming productivity and code performance.

7.1 Benchmarks

The benchmarks we use are based on text analytics in public benchmark suites, HiBench [12] and Puma [3].

Word Count This program counts word frequencies as Section 2.2 already mentions. Its Zwift code has been shown in Listing 2 and discussed in Section 5.

Inverted Index This program identifies, for each word, the set of files containing it, as described in Section 3. We show its Zwift code in Listing 4. It contains two implementations, `Action[0]` and `Action[1]`, and there are two `FinalStages` related to these two different traversals. The Zwift code, equipped with a version selection module created through the offline profiling-based method, automatically selects one of the versions to use for a given input dataset.

Sequence Count This program counts, for each file, the frequencies of each distinct three-word sequence it contains. It has been explained in Section 3. The graph traversal order is DEPTHFIRST.

Ranked Inverted Index This program produces a list of three-word sequences in decreasing order of their occurrences in each document. Similar to `sequence count`, it is also order-sensitive. Its direction is hence set to `DEPTHFIRST`. Its `FinalStage` contains a step to rank all sequences based on their counts.

Sort This program sorts words in alphabetical order. In the Zwift code, the words are stored in the abstract data structure `ZwiftVec`. Its `FinalStage` calls a sorting function on the vector.

Listing 4: Zwift DSL inverted index example.

```
ELEMENT = WORD
USING_FILE = true
EDGE_COARSENING = true
NodeStructure = {
  ZwiftSet<ELEMENT> wordSet;
  ZwiftSet <RULEID> ruleSet;
  ZwiftVec <FILEID,ZwiftBit> file;
}
Init = {
  if(RULEID ==ROOT) {
    for(int i=0; i<FILES; i++){
      int start=getPartition(i);
      int end=getPartition(i+1);
      for(int j=start; j<end; j++){
       if(Zwift_IsWord( ZwiftDAG.Nodes[0].input[j]))
        result[ZwiftDAG.Nodes[0].input[j]].insert(i);
       else
        ZwiftDAG.Nodes[j].file[i]=1;
      }
    }
  }
  else {
    for(vector<int>::iterator i=NODE.input.begin();
       i!=NODE.input.end(); i++) {
      if( Zwift_IsWord(*i) )
        NODE.wordSet.insert(*i);
      else
        NODE.ruleSet.insert(*i);
    }
  }
}
ZwiftNodesExceptRoot = true
Direction[0] = FORWARDS
Direction[1] = BACKWARDS
Action[0] = {
  for(ZwiftVec <FILEID,ZwiftBit>::iterator i=
      PARENT.file.begin();
      i!= PARENT.file.end(); i++){
    CHILD.file[*i]=1;
  }
}
Action[1] = {
  for(ZwiftSet <ELEMENT>::iterator i=
       CHILD.wordSet.begin();
     i!= CHILD.wordSet.end(); i++){
    PARENT.wordSet.insert(*i);
  }
}
Result = {
  ZwiftMap<ELEMENT, ZwiftSet<FILEID> > result;
}
FinalStage[0] = {
  for(int i=1; i<RULES; i++){
    for(ZwiftSet<ELEMENT>::iterator j=
          ZwiftDAG.Nodes[i].wordSet.begin();
        j != ZwiftDAG.Nodes[i].wordSet.end(); j++){
      for(ZwiftVec <FILEID,ZwiftBit>::iterator k=
            ZwiftDAG.Nodes[i].file.begin();
          k != ZwiftDAG.Nodes[i].file.end(); k++)
        if(ZwiftDAG.Nodes[i].file[*k]==1)
          result[*j].insert(*k);
    }
  }
}
FinalStage[1] = {
  for(int i=0; i<FILES; i++){
    int start=getPartition(i);
    int end=getPartition(i+1);
    for(int j=start; j<end; j++){
      if(!Zwift_IsWord(ZwiftDAG.Nodes[0].input[j]))
        for(ZwiftSet <RULEID>::iterator k=
            ZwiftDAG.Nodes[j].wordSet.begin();
            k != ZwiftDAG.Nodes[j].wordSet.end();
            k++)
          result[*k].insert(i);
    }
  }
}
```

Term Vector This program outputs the most frequent words in each file. It needs to keep the file information, so one strategy is to make `ZwiftDAG` transmit the file information from the root to the other nodes first, and then performs `word count` based on files. However, this strategy needs to maintain a copy of word counts for each file in each node, which cost a lot when there are many files. An alternative strategy is to transmit word counts from children to parents in a backward order but do not include the root node by using `ZwiftNodesExceptRoot`. After all nodes finish the "Action", using the root node, which contains file information, and its direct child nodes can easily calculate the word counts for each file and then generate the most frequent words. The code for this version shares some similarity with the backward version of `inverted index`. It can contain both a backward and a forward version.

7.2 Programming Productivity

We measure the number of lines for writing these codes in Table 2. The "C++" column shows the length of the code written for TADOC without using Zwift framework; it employs existing libraries (e.g., C++11 Library) that may help shorten the code. The "Zwift DSL" column shows the length of the equivalent Zwift code. On average, the Zwift code is 84.3% shorter than the equivalent C++ code. Such a comparison is based on only the sequential code. If we count the parallel and distributed versions, the savings are even greater, because the sequential version of Zwift code is sufficient for the Zwift compiler to generate the parallel and distributed versions, while without Zwift, users would need to write all three versions separately. In addition, as Zwift programmer does not need to worry about the graph traversal implementation and the various optimizations related with efficiency, our experience is that the average time per line of code of Zwift is shorter than that of the C++ code. So overall, the time reduction for code development is even greater than the reduction in the length of code.

Table 2: Comparison of code lines between C++ and Zwift.

Programs	C++	Zwift DSL
Word Count	267	70
Sort	279	84
Inverted Index	616	78
Term Vector	719	87
Sequence Count	451	28
Ranked Inverted Index	440	31

7.3 Performance

We measure the performance of Zwift on two platforms, one is a single node machine, and the other is a ten-node cluster on Amazon EC2, as shown in Table 3. We test small datasets on single node machine with Zwift sequential version. For large datasets, we use the ten-node Spark cluster with HDFS storage system [7] to measure Zwift distributed platform.

We use the following datasets in Table 4 in our experiments. `Dataset A` is NSF Research Award Abstracts (NSFRAA) from UCI Machine Learning Repository [15]. This dataset consists of a large number of small files (134,631), and we can use this dataset to measure Zwift's ability of utilizing duplicated information among small files. `Dataset B` to `E` are collections of web documents from Wikipedia database [1]. These datasets consist of large files.

Table 3: Platform Configurations.

Platform	Spark Cluster	Single Node
Operating System	Ubuntu 16.04.1	Ubuntu 14.04.2
Compiler	GCC 5.4.0	GCC 4.8.2
Number of Nodes	10	1
CPU	Intel E5-2676v3	Intel i7-4790
Cores/Machine	2	4
Frequency	2.40GHz	3.60GHz
MemorySize/Machine	8GB	16GB

Table 4: Tested datasets. The size is for the original datasets.

Dataset	File#	Size	Rule#	VocabularySize
dataset A	134,631	580MB	2,771,880	1,864,902
dataset B	4	2.1GB	2,095,573	6,370,437
dataset C	109	50GB	57,394,616	99,239,057
dataset D	309	150GB	160,891,324	102,552,660
dataset E	618	300GB	321,935,239	102,552,660

We measure the space and time benefits in the following two sections. We test the performance of `datasets A` and B on single node and `datasets C`, `D` and `E` on the cluster. We have implemented four versions: (1) *manual-direct*, which is the version on uncompressed data; (2) *manual-gzip*, which is the version that uses Gzip to compress data for space savings, and decompresses the data at the processing time; (3) *manual-opt*, which is the version we manually developed to apply TADOC on our Sequitur compressed data with all the optimizations we have mentioned in the paper except the version selection; in the case of multiple possible directions to implement, it just uses the one working best for most of our testing datasets; (4) *Zwift*, which is the code using our DSL with all the optimizations and adaptations. The baseline is the *manual-direct* version.

7.3.1 Space Benefits. Table 5 reports the space savings by *manual-gzip* and Zwift. Compared to the original uncompressed datasets, Zwift reduces storage usage by 90.8%, even more than *manual-gzip* does. The reason is that Sequitur-based algorithm provides the first round compression, which makes the data smaller than the original size, and then the second-round compression by Gzip makes the data even more compact. Moreover, Table 5 also shows the trend that larger datasets have more potentials for higher compression benefits.

Table 5: Space savings.

Dataset	manual-gzip(%)	Zwift(%)
dataset A	74.66%	77.24%
dataset B	92.14%	94.14%
dataset C	92.40%	95.00%
dataset D	94.40%	96.27%
dataset E	94.33%	96.33%

7.3.2 Time Benefits. We use speedup over $Time_{original}$ as the metric for time benefits. The *manual-gzip* version suffers from the decompression time. Even if the decompression is put to a separate threads to run in parallel with the processing threads, that version cannot run faster than the *manual-direct* version for it adds extra operations while saving none. So we focus the discussions on the comparisons among the other versions.

We show the time benefits of *manual-opt* and Zwift in Figure 4. Zwift yields 2X speedup on average over *manual-direct*, thanks to its reuse of processing results throughout the DAG. Some applications on large datasets, such as `term vector` on `dataset C`, show even more than 70% performance benefits. The reason is that the resilient distributed dataset (RDD), which is the basic storage unit in Spark (recall that the experiments on the large datasets use the Spark version on clusters), has a size limit. Some of the original files exceed the RDD limit and have to be split into separate RDDs. For programs that need file information, such as `inverted index` and `term vector`, additional data structures are needed to be created and maintained to record file information, which adds the cost. Zwift compression decreases file size and most of these large files can fit into the size of RDD after compression, and hence avoid the cost.

From Figure 4, we can see that the time benefits from Zwift are nearly the same as what *manual-opt* gives in almost all cases, which indicates that Zwift successfully unleashes most power of TADOC. The only exceptions are on dataset B on `inverted index` and `term vector`, in which cases, the benefits from Zwift are even larger than *manual-opt* provides. The explanations are two-fold. First, Zwift successfully applies all the optimizations that *manual-opt* does without adding noticeable extra overhead. Second, the version selection by Zwift exhibits performance benefits, making it outperform *manual-opt*. Recall that *manual-opt* does not do version selection. Instead, it implements backward traversal for `inverted index` and `term vector` as it works the best on them on most datasets. Dataset `B` is an exception as it contains only four large files. The forward version has a lower cost for both benchmarks. Zwift's cost model helps it select the right version to run.

The offline profiling of version selection of Zwift costs negligible overhead. On `inverted index` on dataset B, for instance, the overhead is less than one millisecond, while the program runs for 13.2 seconds.

Zwift data abstractions turn out to be quite helpful. `ZwiftVec` brings in significant space benefits. For example, in dataset `A`, Zwift saves 99.4% space through the usage of two-level bit vector for instantiating `ZwiftVec`, compared to the use of normal `vector` data structure. The reduced memory footprint contributes to better cache and TLB performance. The coarsening of `ZwiftDAG` at loading time gives good benefits as well. Edge coarsening is enabled for all benchmarks except `sequence count` and `ranked inverted index` for their order sensitivity. Node coarsening is enabled for `sequence count` and `ranked inverted index` benchmarks. On dataset `B`, for instance, node coarsening (with a threshold of 100) helps save 5% time over the runs without node coarsening.

7.4 Discussions

This part discusses the time taken to compress datasets, the implications, and other aspects of the applicability of Zwift.

In our experiments, a dataset takes Sequitur 10 minutes to 20 hours to compress, depending on the dataset size. The current Zwift

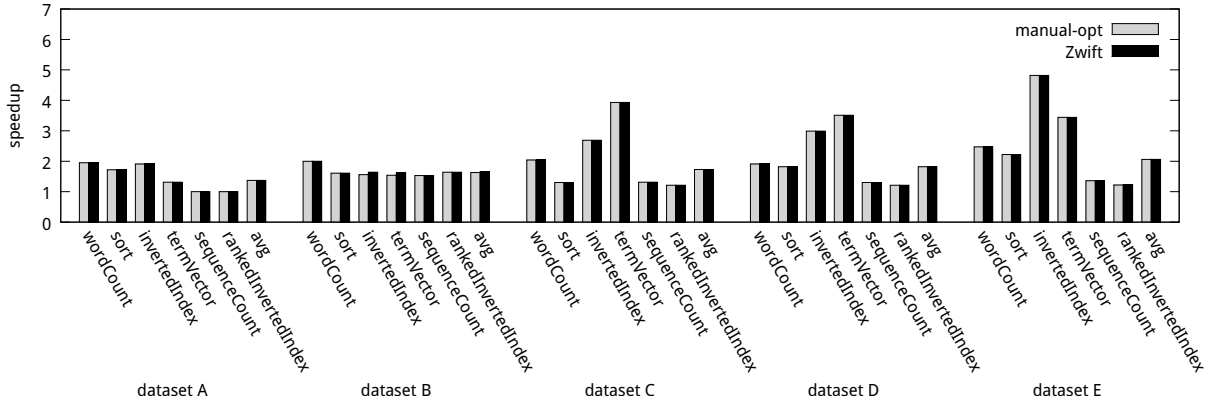

Figure 4: Time benefits.

is designed for datasets that are repeatedly used without being frequently updated. For them, the compression result of a dataset can be used many times by different users for various tasks, and the one-time compression time is not a major concern. To shorten the time, one may employ more efficient parallel or distributed implementations of Sequitur. How to extend Zwift to handle datasets with frequent updates is for future explorations.

Zwift mainly focuses on applications that normally require scanning the entire dataset. Another type of common tasks involve queries that require random accesses to some locations in the dataset. They are much simpler to support; adding some appropriate index to the Sequitur results could do the job. Such tasks can already be supported by other recent techniques (e.g., Succinct [2]), and are hence not included in the focus of this paper.

8 RELATED WORK

There are many frameworks or languages for specific domains [4, 5, 9, 13, 22, 26, 32]. Some examples include PetaBricks [4] designed for algorithmic tuning, OpenTuner [5] for building program autotuners, Halide [22] for image processing pipeline, HPTA [26] for high-performance text analytics, and TCS [13] for model-to-text and text-to-model transformations. Zwift is the first framework designed for text analytics directly on compressed data.

Deduplication is a topic actively studied in storage systems. Hernández and others [11] pointed out the data merge and purge problem. String matching [17] and file matching [16] have been studied for deduplications. Sarawagi and others [24] developed a learning-based deduplication system that focused on multiple sources. Raman et al. [23] developed an interactive data cleaning system that enabled users build transformations to clean data. As an OS-level technique, data deduplication does not save computations in text analytics as TADOC does. If two files contain overlapped content, even if deduplication manages to save only one copy of the overlapped content, a text analytics application still processes the content repeatedly for the two files.

Succinct [2] enables efficient queries on databases with data compression. It is specific to database domain, supporting searching and random accessing of a specific query only. TADOC supports much more complex document analytics tasks. For example, none of the six analytics problems used in our experiments are amenable for efficient support by Succinct, as they involve complex algorithms and operations beyond search or random access. Moreover, they use different compression methods and inner storage structures. Succinct uses a flat structure, suffix array [18], while Zwift uses Sequitur to create a DAG structure, allowing it to perform complex computations for all items.

9 CONCLUSION

This paper introduces Zwift, a new programming framework to enable general programmers to take advantage of TADOC (text analytics directly on compressed data) to save both storage space and time in high-performance text analytics. By relieving programmers from most of the efficiency concerns and complexities for implementing text analytics on compressed data, Zwift significantly reduces the efforts required for general programmers to apply TADOC while still unleashing its full power for efficient text analytics. We conclude that Zwift is an effective framework for general-purpose TADOC, and hope that future research builds on it to demonstrate benefit on even larger data analytics problems.

10 ACKNOWLEDGEMENTS

This work is partially supported by National Key R&D Program of China (Grant No. 2016YFB0200100), National Natural Science Foundation of China (Grant No. 61722208, 61472201, 61732014), Tsinghua University Initiative Scientific Research Program (20151080407). This material is based upon work supported by DOE Early Career Award (DE-SC0013700), the National Science Foundation (NSF) under Grant No. CCF-1455404, CCF-1525609, CNS-1717425, CCF-1703487. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DOE or NSF. This work is also supported by the Beijing Natural Science Foundation (No. 4172031), China Postdoctoral Science Foundation (2017M620992), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 16XNLQ02, 18XNLG07), and the open research program of State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Science (No. CARCH201702). Jidong Zhai and Xipeng Shen are the corresponding authors of this paper.

REFERENCES

[1] 2017. Wikipedia HTML data dumps. https://dumps.wikimedia.org/enwiki/. (2017).

[2] Rachit Agarwal, Anurag Khandelwal, and Ion Stoica. 2015. Succinct: Enabling queries on compressed data. In *NSDI*. 337–350.

[3] Faraz Ahmad, Seyong Lee, Mithuna Thottethodi, and TN Vijaykumar. 2012. Puma: Purdue mapreduce benchmarks suite. (2012).

[4] Jason Ansel, Cy Chan, Yee Lok Wong, Marek Olszewski, Qin Zhao, Alan Edelman, and Saman Amarasinghe. 2009. *PetaBricks: a language and compiler for algorithmic choice*. Vol. 44. ACM.

[5] Jason Ansel, Shoaib Kamil, Kalyan Veeramachaneni, Jonathan Ragan-Kelley, Jeffrey Bosboom, Una-May O'Reilly, and Saman Amarasinghe. 2014. Opentuner: An extensible framework for program autotuning. In *PACT*. ACM, 303–315.

[6] Joshua E Blumenstock. 2008. Size matters: word count as a measure of quality on wikipedia. In *WWW*. ACM, 1095–1096.

[7] Dhruba Borthakur. 2008. HDFS architecture guide. *HADOOP APACHE PROJECT http://hadoop. apache. org/common/docs/current/hdfs design. pdf* (2008), 39.

[8] Doug Cutting and Jan Pedersen. 1989. Optimization for dynamic inverted index maintenance. In *SIGIR*. ACM, 405–411.

[9] Dapeng Dong and John Herbert. 2017. Content-aware Partial Compression for Textual Big Data Analysis in Hadoop. *IEEE Transactions on Big Data* (2017).

[10] Wenting He, Huimin Cui, Binbin Lu, Jiacheng Zhao, Shengmei Li, Gong Ruan, Jingling Xue, Xiaobing Feng, Wensen Yang, and Youliang Yan. 2015. Hadoop+: Modeling and evaluating the heterogeneity for mapreduce applications in heterogeneous clusters. In *ICS*. ACM, 143–153.

[11] Mauricio A Hernández and Salvatore J Stolfo. 1998. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data mining and knowledge discovery* 2, 1 (1998), 9–37.

[12] Shengsheng Huang, Jie Huang, Jinquan Dai, Tao Xie, and Bo Huang. 2011. The HiBench benchmark suite: Characterization of the MapReduce-based data analysis. In *New Frontiers in Information and Software as Services*. Springer, 209–228.

[13] Frédéric Jouault, Jean Bézivin, and Ivan Kurtev. 2006. TCS: a DSL for the specification of textual concrete syntaxes in model engineering. In *Proceedings of the 5th international conference on Generative programming and component engineering*. ACM, 249–254.

[14] Ludovic Lebart. 1998. Classification problems in text analysis and information retrieval. In *Advances in Data Science and Classification*. Springer, 465–472.

[15] M. Lichman. 2013. UCI Machine Learning Repository. (2013). http://archive.ics.uci.edu/ml

[16] Alvaro E Monge, Charles Elkan, et al. 1996. The Field Matching Problem: Algorithms and Applications. In *KDD*. 267–270.

[17] Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM computing surveys (CSUR)* 33, 1 (2001), 31–88.

[18] Gonzalo Navarro. 2016. *Compact Data Structures: A Practical Approach*. Cambridge University Press.

[19] Craig G. Nevill-Manning and Ian H. Witten. 1997. Identifying hierarchical structure in sequences: A linear-time algorithm. *Journal of Artificial Intelligence Research* 7 (1997), 67–82.

[20] Nicholas A Nystrom, Michael J Levine, Ralph Z Roskies, and J Scott. 2015. Bridges: a uniquely flexible HPC resource for new communities and data analytics. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*. ACM, 30.

[21] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001), 2001.

[22] Jonathan Ragan-Kelley, Connelly Barnes, Andrew Adams, Sylvain Paris, Frédo Durand, and Saman Amarasinghe. 2013. Halide: a language and compiler for optimizing parallelism, locality, and recomputation in image processing pipelines. *ACM SIGPLAN Notices* 48, 6 (2013), 519–530.

[23] Vijayshankar Raman and Joseph M Hellerstein. 2001. Potter's wheel: An interactive data cleaning system. In *VLDB*, Vol. 1. 381–390.

[24] Sunita Sarawagi and Anuradha Bhamidipaty. 2002. Interactive deduplication using active learning. In *SIGKDD*. ACM, 269–278.

[25] Julian Shun and Guy E Blelloch. 2013. Ligra: a lightweight graph processing framework for shared memory. In *ACM Sigplan Notices*, Vol. 48. ACM, 135–146.

[26] Hans Vandierendonck, Karen Murphy, Mahwish Arif, and Dimitrios S Nikolopoulos. 2016. HPTA: High-performance text analytics. In *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 416–423.

[27] Hao Yan, Shuai Ding, and Torsten Suel. 2009. Inverted index compression and query processing with optimized document ordering. In *WWW*. ACM, 401–410.

[28] Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. 2010. Spark: cluster computing with working sets. *HotCloud* 10 (2010), 10–10.

[29] Uri Zernik. 1991. *Lexical acquisition: exploiting on-line resources to build a lexicon*. Psychology Press.

[30] Feng Zhang, Jidong Zhai, Xipeng Shen, and Onur Mutlu. 2017. *Potential of A Method for Text Analytics Directly on Compressed Data*. Technical Report TR-2017-4. NCSU.

[31] Fang Zheng, Hasan Abbasi, Ciprian Docan, Jay Lofstead, Qing Liu, Scott Klasky, Manish Parashar, Norbert Podhorszki, Karsten Schwan, and Matthew Wolf. 2010. PreDatA–preparatory data analytics on peta-scale machines. In *IPDPS*. IEEE, 1–12.

[32] Zhen Zheng, Chanyoung Oh, Jidong Zhai, Xipeng Shen, Youngmin Yi, and Wenguang Chen. 2017. Versapipe: a versatile programming framework for pipelined computing on GPU. In *MICRO*. ACM, 587–599.

[33] Jacob Ziv and Abraham Lempel. 1977. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory* 23, 3 (1977), 337–343.