10장. 함수형 데이터 구조

데이터 구조는 알고리즘의 핵심 요소이다. 힙, 큐, 스택, 트리, 그리고 해시 테이블은 프로그래밍 언어에서 널리 사용되는 데이터 구조의 여러 형태이다. 그 중 일부는 트리나 해시 테이블과 같이 주로 검색에 사용되고, 힙, 큐, 스택과 같은 나머지는 삽입 및 삭제 등의 업데이트를 위해 사용된다. 10장에서는 평범한 데이터 구조를 함수형 데이터 구조로 확장하기 위한 기초를 수립할 것이다. 10장에서 배울 개념은 다음과 같다.

* 함수형 데이터 구조
* 지연된 평가(lazy evaluation)
* 함수형 큐
* 함수형 스택

<대> 함수형 데이터 구조

함수형 데이터 구조는 주로 함수형 프로그래밍 언어에서 구현되는 데이터 구조의 특별한 형태이다. R은 함수의 생성 및 처리를 위한 도구를 제공함으로써 함수형 프로그래밍을 지원한다. 예를 들어, R은 함수를 변수에 할당하고 그것을 함수 내에서 인수로 전달하는 것을 지원한다. R은 동적으로 함수를 생성하고 함수의 결과로서 그것을 반환하는 클로저 함수(closure function)를 지원한다. 예를 들어, 함수를 인수로 취하는 함수는 다음과 같다.

arg\_function <- function(g) g(seq(1, 100, by=1))

arg\_function 함수는 mean 이나 sd 같은 함수를 인수로 취하며, 그 코드 예제는 다음과 같다.

> arg\_function(mean)

[1] 50.5

> arg\_function(sd)

[1] 29.01149

함수형 데이터 구조는 지속성 데이터 구조라고도 한다. 함수형 데이터 구조에 대해 수행된 모든 작업은 업데이트 된 작업과 함께 데이터 구조의 새로운 복사본을 생성한다는 의미에서 불변이기 때문이다. 함수형 데이터 구조의 원본은 항상 그대로 유지된다. 함수형 데이터 구조는 전통적인 데이터 구조와는 다른 특성을 가지고 있다. 함수형 데이터 구조는 구현시에 매우 유연하며 불변성(immutability)과 지속성(persistency)을 지원한다. 또한 쓰레드 안정성(thread-safe)도 지원한다. 이는 멀티쓰레드 환경에서 데이터 구조를 처리할 때 안정적인 실행을 보장한다.

다음은 불변성이 주는 이점들이다.

* 데이터 은닉 및 데이터 공유를 지원한다. 데이터 은닉은 데이터(함수 내에서 동적으로 생성됨)의 유출 가능성을 방지하고, 데이터 공유는 함수 내에서 필요에 따른 데이터의 자동 공유를 지원한다(포인터 기반).
* 함수형 데이터 구조는 변형을 지원하지 않기 때문에 외부 동기화를 할 필요가 없다. 그러나 통제된 변형을 통한 수정(추가 또는 삭제)을 일부 지원한다.
* 포인터가 장거리 연결되는 것을 방지한다.

다음은 지속성이 주는 이점들이다.

* 파괴적인 업데이트 대신에 데이터 구조에 대한 건설적인 업데이트를 통해 변형이 일어나도록 한다. 수정은 이전 버전은 그대로 유지하고 전체 데이터 구조가 복제되는 것으로 통합된다.
* 단순한 복사가 아닌 이전 버전의 공유와 같은 메모리 효율적인 방식의 복제를 지원한다.

다음은 쓰레드 안정성이 주는 이점들이다.

* 데이터 변형에 대한 걱정 없이 동시성 프로그래밍을 지원한다.
* 새로운 객체를 생성하는 동안, 변형은 포인터가 가리키는 쓰레드 내에서만 수행되며, 원본 데이터는 그대로 유지된다(불변).
* 공유 포인터를 사용하는 쓰레드 안전 참조 카운팅을 지원하고, 로그 프리(log-free)로 최적화 되기 때문에 메모리 효율적이다.

<중> 지연된 평가

함수형 프로그래밍 언어에서 사용되는 함수형 데이터 구조는 지연된 평가를 지원한다. 함수의 인수는 계산 결과가 함수 내에서 더 사용될 경우에만 평가된다. 더 나아가 인수가 평가되고 나면 계산 결과는 캐시(cache)로 저장되고, 다시 필요한 경우 재계산을 하는 대신에 저장된 결과를 재사용한다. 이런 유형의 캐싱(메모이제이션 memoization 이라고도 한다)은 주어진 인수(또는 부분 표현식)가 언제 평가되는지 결정하기 쉽지 않기 때문에 알고리즘의 점근적 복잡도를 추정하기 아주 어렵게 만든다.

지연된 평가는 함수형 데이터 구조를 구현할 때 핵심적인 역할을 한다. 지연된 평가가 포함된 알고리즘의 점근적 성능을 분석하는 것은 매우 어렵다. 하지만 다음 프레임워크는 지연된 평가가 포함된 알고리즘의 점근적 성능을 계산하기 위한 기초를 제공한다. 우선, 작업의 비용은 다음과 같이 분류된다.

* 비공유 비용(Unshared cost) : 작업의 실제 실행 시간으로 정의된다. 작업을 실행하기 전에 이 작업과 관련되어 연기된 모든 작업이 강제되고 메모된다.
* 공유 비용(Shared cost) : 현재 실행중인 작업의 성능을 평가하느라 이전에 제외되었던 모든 연기된 작업의 실행 시간으로 정의된다. 공유 비용은 실현 비용과 미실현 비용으로 더 나눌 수 있다. 실현 비용은 전체 연산 중에 실행되는 연기된 작업의 런타임을 평가하는 반면, 미실현 비용은 전체 연산 중에 전혀 실행되지 않는 연기된 작업의 런타임을 평가한다.
* 전체 비용(Complete cost) : 지연된 평가를 사용하여 구현된 작업의 실제 실행 시간으로 정의된다. 이것은 미실현 비용을 제외한 공유 비용과 비공유 비용의 합이다. 전체 연산 중에 실행되는 지연된 작업이 없으면 전체 비용의 최소값은 비공유 비용과 같다.
* 상각 비용(Amortized cost) : 누적 채무의 개념을 사용하여 계산된 전체 연산의 공유 비용으로 정의된다. 처음에 누적 채무는 0으로 설정되고, 지연이 발생할 때마다 누적된다. 그 다음에 각 작업은 시작될 때 채무 상환을 한다. 채무가 완전히 상환되면 작업에 대한 지연이 강제되고 메모될 수 있다. 여기서 작업의 상각 비용은 작업의 비공유 비용과 각 작업에 의해 지불된 채무의 합이다.

따라서 작업 연기는 누적된 채무가 완전히 상환될 때만 강제되고 메모이제이션이 될 수 있다. 누적 비용의 총계가 실현된 공유 비용으로 제한되기 때문에 상각 비용은 실제 총 비용을 넘어설 수 없다.

<중> 함수형 스택

스택은 선입후출(First In Last Out, FILO) 형식의 데이터 구조로, 삽입과 삭제는 일반적으로 스택의 최상단(top)에서 발생한다. 이 형식은 깊이 우선 탐색(DFS)에서 널리 사용된다. 이상적인 시나리오에서 삽입, 삭제 및 검색 작업은 매우 적은 시간이 요구되며, 일반적으로 최악의 경우 시나리오에 대해서 O(1)의 점근선을 갖는다. 따라서 최악의 경우 시나리오에서 n 개의 개체 삽입 또는 삭제는 O(n)의 점근선을 가지며, 각 작업의 평균 점근선은 O(1)이다.

이제 R에서 완전 지속성 스택(fully-persistent stack)의 구현에 대해 알아보자. 완전 지속성 스택 데이터 구조의 구현은 숀 오네일(Shawn T. O'Neil)이 제공한 rstackdeque 패키지를 사용하면 된다.

이 패키지에서 스택은 정렬되지 않은 링크드 리스트를 사용하여 구현되며, 각 노드(리스트)는 데이터 개체와 다음 노드에 대한 참조로 구성되어 있다. 이 스택은 S3 객체로 스택의 헤드 노드를 통해 액세스 가능하다.

5 개의 문자 개체를 가진 스택이 있다.

a <- as.rstack(c("p", "q", "r","s","t"))

일반적인 push 함수와 비슷하게 insert\_top은 새 개체가 최상단에 추가된 스택을 반환하는데 사용된다. 여기서 새로운 스택을 생성하는 대신에, 그림 10.1에 보이는 것처럼 스택 b의 헤드는 o 개체를 가리키고, o 개체는 차례로 스택 a를 가리킨다.

b <- insert\_top(a, c("o"))

스택에서 최상단 개체를 삭제(pop)하는 경우에는 without\_top 함수가 사용된다. insert\_top과 마찬가지로 기본 스택이 파괴적으로 업데이트 되는 것이 아니라 스택 c의 포인터는 그림 10.1에 보이는 것처럼 한 개체 오른쪽으로 이동한 위치를 가리킨다.

c <- without\_top(a)

peek\_top 함수는 스택의 최상단에 있는 데이터 개체를 반환할 때 사용된다.

d <- peek\_top(a)

그림 10.1은 R에서 완전 지속성 스택의 구현을 도식화한 것이다.



<그림 10.1: 삽입 또는 삭제 작업에 대한 완전 지속성 스택의 동작. (a) 5 개의 문자를 가진 스택을 나타낸다. 각 문자는 다른 문자와 링크드 리스트로 연결되어 있다. (b) 스택 a의 최상단에 새 개체 “o”를 추가한 후의 스택. (c) 스택 a에서 최상단 개체를 삭제한 후의 스택. (d) 스택 b의 최상단 개체를 가진 문자 벡터>

(## 원서 오류 : top element of stack (a) -> stack (b). stack (a)의 최상단 개체는 “o”가 아닌 “p”이다.)

<중> 함수형 큐

큐는 선입선출(First In First Out, FIFO) 형태의 데이터 구조로 삽입된 순서와 동일한 순서대로 삭제가 발생한다. 큐를 구현하는 한 가지 방법은 후단(큐의 최상단)에서 개체를 삽입하고 반대편 끝인 전단(큐의 최하단)에서 개체를 삭제하는 것이다. 일부 큐는 양쪽 끝 모두에서 삽입과 삭제를 지원한다. 이를 덱(deque) 또는 양방향 큐(double-ended queue)라고 부른다. 큐와 덱은 너비 우선 탐색(BFS) 알고리즘에서 사용된다. 스택과 마찬가지로 이상적인 삽입, 삭제 및 검색 작업은 최악의 경우 시나리오에 대해 O(1)의 점근선으로 매우 적은 시간을 필요로 한다. 따라서 최악의 경우 시나리오에서 n 개의 삽입 또는 삭제는 O(n)의 점근선을 가지며, 각 작업의 평균 점근선은 O(1)이다.

이제 rstackdeque 패키지를 사용하여 빠른 완전 지속성, 그리고 느린 지속성 큐와 덱을 구현해보자.

<소> 빠른 완전 지속성 큐

빠른 완전 지속성 큐(fast fully-persistent queue)는 주로 재귀적으로 정의된 작업과 지연된 평가에 의해 관리된다. 즉, 이 큐는 첫 번째 개체는 즉시 액세스할 수 있고 나머지 개체는 지연 시간을 갖고 평가되는 지연 리스트(lazy list)를 사용하여 구현된다. 개체는 오직 액세스될 때만 평가되기 때문에 이것은 대규모의 리스트를 반복적으로 처리하는 경우에 유용하다.

빠른 지속성 큐를 구현하는 R 함수는 rpqueue이다. 이것은 l, r, lhat 이란 세 개의 리스트로 구성되어 있다. 여기서 개체는 큐의 후단에서 삽입되고, 큐의 전단에서 삭제된다. 즉, 새 개체의 삽입은 r 리스트의 최상단에서 일어나고, 기존 개체의 삭제는 l 리스트의 전단에서 일어난다. 이 지연 리스트는 rstacks으로 구현되며, 모든 노드의 nextnode는 delayedAssign 함수를 통해 할당된다. 첫 번째 평가 후에 이것들은 메모된다.

일반적으로 큐는 l 스택의 길이를 최소한 r 스택의 길이와 동일하게 유지하려고 한다. 그러나 삽입 또는 삭제 작업 동안에는 그럴 수 없게 된다. 삽입 작업에 의해 r 스택의 길이는 증가하고, 삭제 작업에 의해 l 스택의 길이는 감소한다. 이 작업들 후에 r 스택의 마지막 개체를 l 스택에 추가함으로써 l 스택과 r 스택은 다시 조정된다. 이러한 재조정은 l 스택과 r 스택 내에서 발생하는 삽입 또는 삭제에 따라 O(1)의 점근선을 필요로 한다.

Traditionally, the queue tries to ensure that the length of l stack is at least equal to the length of the r stack. However, during any insertion or deletion operation, the lengths are disturbed. The length of r stack increases upon insertion and the length of l stack decreases upon deletion. Post these operations, the l and r stacks are readjusted by appending the last elements of r stack with l stack. This readjustment requires an asymptote of O(1) for any sort of insertion or deletion happening within l and r stacks.

재조정에 따라 lhat에 l 스택의 데이터가 할당된다. 그 다음에 차후의 각 삽입 또는 삭제 작업에서 lhat의 개체는 하나씩 제거된다. lhat이 비게 되면 l 스택과 r 스택의 길이가 평가되고 개체들은 다시 재조정되어 반복 시간이 지연된다.

Upon readjustment, lhat is assigned with data of stack l . Then, for each subsequent insertion or deletion, the elements of lhat are removed one by one until the lhat stack becomes empty. Once lhat becomes empty, the lengths of l and r stacks are evaluated and the elements are readjusted, again delaying the time of iteration.

예제를 사용하여 완전 지속성 큐의 동작을 이해해보자. 길이가 4인 지속성 큐가 있다.

Let's understand the working of fully-persistent queues using an example. Consider a persistent queue of length four:

a <- as.rpqueue(c("p","q","r","s"))

왼쪽 스택 l에는 r, q, r 개체가 할당되고, 오른쪽 스택 r에는 s 개체가 할당되며, lhat 스택에는 q와 r이 할당된다. 다음 그림 10.2는 삽입과 삭제 작업에 따른 지속성 큐의 동작을 보여준다. 삽입 작업은 insert\_back 함수를 사용하여 수행되며, 삭제 작업은 without\_front 함수를 통해 수행된다.

The "p" , "q" , and "r" elements are assigned to left stack l , element "s" is assigned to right stack "r" , and, elements "q" and "r" are assigned to stack lhat . The following Figure 10.2 illustrates working of persistent queues based on insertion and deletion operations. The insertion operation is performed using the insert\_back function and deletion operation is performed using the without\_front function.

<그림시작>

새로운 지속성 큐 생성.

큐의 후단에 새 개체 “t”를 삽입. 이 작업은 lhat에서 한 개체를 삭제한다. 하지만 lhat이 아직 비지 않았기 때문에 l과 r 스택의 재조정은 일어나지 않는다.

큐의 전단에서 개체 “p” 삭제. 다시 이 작업은 lhat에서 한 개체를 삭제한다. 하지만 lhat이 아직 비지 않았기 때문에 l과 r 스택의 재조정은 일어나지 않는다.

큐의 후단에 새 개체 “v”를 삽입. lhat이 비어 있기 때문에 l과 r 스택 사이에서 재조정이 발생한다. 모든 개체가 왼쪽 스택 l과 lhat에 할당된다. 이 작업은 재조정으로 인해 런타임을 지연시킨다.

큐의 후단(r 스택)에 새 개체 “w”를 삽입. 이 작업은 lhat에서 한 개체를 삭제한다. 하지만 lhat이 아직 비지 않았기 때문에 l과 r 스택의 재조정은 일어나지 않는다.

큐의 전단에서 개체 “q” 삭제. 다시 이 작업은 lhat에서 한 개체를 삭제한다. 하지만 lhat이 아직 비지 않았기 때문에 l과 r 스택의 재조정은 일어나지 않는다.

<그림끝>

<그림 10.2: 삽입 또는 삭제 작업에 따른 완전 지속성 큐의 동작. (a) 새로운 지속성 큐. (b) 큐 a의 후단에 새 개체 “t”를 삽입한 후의 큐. (c) 큐 b에서 전단에 있는 개체를 삭제한 후의 큐. (d) 큐 c의 후단에 새 개체 “v”를 삽입한 후의 큐. (e) 큐 d의 후단에 새 개체 “w”를 삽입한 후의 큐. (f) 큐 e에서 전단에 있는 개체를 삭제한 후의 큐.>

Figure 10.2: Working of fully-persistent queues based on diﬀerent types of insertion or deletion operations. (a) represents a new persistent queue. (b) represents a queue after inserting a new element “ t ” in the back of queue (a). (c) represents a queue after deleting the front element from queue (b). (d) represents queue after inserting a new element “ v ” at the back of queue (c). (e) represents queue after inserting a new element “ w ” at the back of queue (d). (f) represents a queue after deleting an element from front of queue (e).

다음은 앞의 그림에서 사용한 R 코드이다.

The following are R codes, which are used in the preceding illustration:

b <- insert\_back(a, "t") # 새 개체 "t" 삽입

c <- without\_front(b) # b에서 전단 개체 삭제

d <- insert\_back(c,"v") # 새 개체 "v" 삽입

e <- insert\_back(d,"w") # 새 개체 "w" 삽입

f <- without\_front(e) # e에서 전단 개체 삭제

<소> 느린 지속성 큐와 덱

큐는 왼쪽 스택과 오른쪽 스택, 두 개의 스택으로 구현된다. 왼쪽 스택은 삭제 작업에 사용되는 큐의 첫 번째 개체 집합을 가지고 있으며, 오른쪽 스택은 삽입 작업에 사용되는 큐의 마지막 개체 집합을 가지고 있다. 다른 한편, 왼쪽 스택 역시 삽입 작업에 사용될 수 있으며, 오른쪽 스택도 삭제 작업에 사용될 수 있다. 오른쪽 스택은 그림 10.3에 보이는 것처럼 개체를 역순으로 가지고 있다.

The queues are implemented as two stacks, that is, the left stack and the right stack. The left stack holds the first set of elements of the queue, which is used for deletion operations, and the right stack holds the last set of elements of the queue, which is used for insertion operations. On the other hand, the left stack can also be used for insertion and the right stack for deletion. The right stack holds elements in the reverse order as shown in Figure 10.3 .

7 개의 문자를 개체로 가진 큐가 있다.

Let's consider an a queue with seven character elements:

a <- as.rdeque(c("p", "q", "r","s","t","u","v"))

큐는 그림 10.3과 10.4에 보이는 것처럼 왼쪽 큐와 오른쪽 큐로 나눠진다. insert\_front 또는 insert\_back 함수를 통해 개체는 전단 또는 후단에 각각 삽입된다.

The a queue is split into left and right queues, as illustrated in Figure 10.3 and Figure 10.4 . Using the insert\_front or insert\_back functions, the elements are inserted at the front or back of the queue respectively:

b <- insert\_front(a, c("o"))

c <- insert\_back(a, c("w"))

마찬가지로 개체는 without\_front 또는 without\_back 함수를 통해 큐의 전단 또는 후단에서 삭제될 수 있다.

Similarly, elements can also be deleted from the front or back of the queue using the without\_front or without\_back functions:

d <- without\_front(a)

e <- without\_back(a)

다음 그림 10.3과 10.4는 R에서 느린 지속성 큐의 구현을 설명한다.

The following Figure 10.3 and Figure 10.4 describes implementation of slowly-persistent queues in R:

<그림 10.3: 삽입 또는 삭제 작업에 따른 느린 지속성 큐의 동작. (a) 왼쪽 스택과 오른쪽 스택으로 구성된 큐. (b) 큐 a의 전단에 새 개체 “o”를 삽입한 후의 큐.>

Figure 10.3: Working of slowly-persistent queues and dequeus based on diﬀerent types of insertion or deletion operations. (a) represents a queue with left and right stacks. (b) represents queue after inserting a new element “ o ” at the front of queue (a).

그림 10.3은 느린 지속성 큐가 처음 7 개의 문자 개체를 어떻게 저장하는지 보여준다(a). 큐의 왼쪽은 a$l 명령을 통해 확인할 수 있으며, 마찬가지로 큐의 오른쪽은 a$r 명령을 통해 내용을 확인할 수 있다. (b)는 insert\_front 함수를 통해 개체 “o”가 추가되었을 때 함수형 큐가 어떻게 업데이트 되는지 보여준다.

The Figure 10.3 presents how slowly-persistent queues stores initial seven character element passed to a . The left side of the queue can be extracted using a$l and similarly right side of queue can be extracted using a$r . The b part of image demonstrate how functional queue a will be updated when an element o is added using function insert\_front(a, c("o")) .

<그림 10.4: (c) 큐 a의 후단에 새 개체 “w”를 삽입한 후의 큐. (d) 큐 a에서 전단 개체를 삭제한 후의 큐. (e) 큐 a에서 마지막 개체를 삭제한 후의 큐.>

Figure 10.4: (c) represents queue after inserting a new element w in the back of queue (a). (d) represents queue after deleting the front element from queue (a). (e) represents queue after deleting the last element from queue (a).

이 구현에서는 그림 10.4에 보이는 것처럼 모든 삽입 또는 삭제 후에 균형을 유지하는 양방향 큐를 사용하였다. 만약 왼쪽과 오른쪽 모두 매우 불균형하게 되면 먼저 둘 다 하나의 리스트로 분해된 다음에 다시 균형이 맞는 두 개의 스택으로 재구성된다.

In the current implementation, double-ended queues, which rebalance after every insertion or deletion, are used as shown in Figure 10.4 . If both left and right stacks become highly unbalanced, then they are both first decomposed into a list and then recomposed back into two nearly balanced stacks.

<대> 요약

함수형 프로그래밍은 학문적인 목적 외에도 항공우주, 통신, 로보틱스와 같은 산업 분야에서 폭넓게 사용된다. 하스켈(Haskell), 스칼라(Scala), F#과 같은 순수한 함수형 언어가 이들 산업에서 유명해졌다. 불변성과 지속성은 투명성, 시뮬레이션, 그리고 효율성 관점에서 이 데이터 구조를 더욱 강력하게 만든다. 10장에서는 함수형 데이터 구조, 지연 평가의 개념, 그리고 함수형 스택과 큐에 대한 자세한 분석을 다뤘다.

Apart from academic purposes, functional programming has found its way into a wide range of industrial applications such as aerospace, telecommunications and robotics. Pure functional languages such as Haskell, Scala, F# are becoming popular across these industries. The properties of immutability and persistency make these data structures more robust in terms of transparency, simulations and efficiency. The current chapter primarily covers functional data structures, concepts of lazy evaluation and detailed analysis of functional stacks and functional queues.