ФЕДЕРАЛЬНОЕ АГЕНТСТВО СВЯЗИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ОБРАЗОВАНИЯ

«СИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ТЕЛЕКОММУНИКАЦИЙ И ИНФОРМАТИКИ»

**ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

к курсовому проекту по дисциплине

«Структуры и алгоритмы обработки данных»

на тему

ТЕРНАРНОЕ ДЕРЕВО ПОИСКА

|  |  |
| --- | --- |
| Выполнил студент | Кулик Павел Евгеньевич |
|  | Ф.И.О. |

|  |  |
| --- | --- |
| Группы | ИС241 |
|  |  |

|  |  |  |
| --- | --- | --- |
| Работу принял |  | ст. преп. Кафедры ВС Д. М. Берлизов |
|  | подпись |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Защищена |  | Оценка |  |
|  |  |  |  |

Новосибирск – 2023

Оглавление

[ВВЕДЕНИЕ 3](#_Toc151331145)

[ОПИСАНИЕ ОСНОВНЫХ ОПЕРАЦИЙ 5](#_Toc151331146)

[1. Вставка ключа в дерево 5](#_Toc151331147)

[2. Удаление ключа из дерева 6](#_Toc151331148)

[2.1 Вспомогательные операции для удаления 8](#_Toc151331149)

[2.1.1 start\_of\_key 8](#_Toc151331150)

[2.1.2 end\_of\_key 9](#_Toc151331151)

[2.1.3 free\_key 10](#_Toc151331152)

[2.1.4 tst\_restore\_prop\_any\_case 10](#_Toc151331153)

[2.1.5 tst\_restore\_prop 11](#_Toc151331154)

[3. Удаление всего дерева 12](#_Toc151331155)

[4. Вывод списка слов 12](#_Toc151331156)

[4.1 tst\_print\_all\_words 12](#_Toc151331157)

[4.2 tst\_print\_words\_from\_node 12](#_Toc151331158)

[5. Поиск 13](#_Toc151331159)

[5.1 tst\_lookup 13](#_Toc151331160)

[5.2 tst\_find\_sym 14](#_Toc151331161)

[6. Префиксный поиск 14](#_Toc151331162)

[6.1 tst\_prefix\_search 14](#_Toc151331163)

[АНАЛИЗ ЭФФЕКТИВНОСТИ АЛГОРИТМОВ 15](#_Toc151331164)

# ВВЕДЕНИЕ

На сегодняшний день разработано и исследовано множество различных структур данных, каждая из которых обладает своими особенностями и может эффективно применяться в подходящих задачах.

В рамках данной курсовой работы мною была изучена и реализована такая структура данных как тернарное дерево поиска. Впервые эта структура была описана в статье под названием "A tree for associative searching" («Дерево для ассоциативного поиска»), написанной Найджелом П. Калдером (Nigel P. Chapman) и Эрнестом С. Коузенсом (Ernest S. Couden). Эта статья была опубликована в журнале "Communications of the ACM" в июне 1971 года (Volume 14, Issue 6, Pages 427-432).

Исследуемая структура может обладать множеством различных полезных применений:

1. Автозавершение и поиск по словарям: TST может быть использовано для реализации функциональности автозавершения в текстовых редакторах, поиске по словарям или поиске слов в больших наборах данных.
2. Индексация текста: тернарные деревья поиска могут быть применены для построения индекса текста, что ускоряет операции поиска подстрок в текстовых данных.
3. Частотный анализ: TST может использоваться для анализа частоты встречаемости слов в тексте. Это полезно в задачах анализа текста и обработки естественного языка.
4. Словари и поиск слов: тернарные деревья поиска могут быть применены для хранения словарей и эффективного поиска слов в них.
5. Телефонные книги и каталоги: TST может быть использовано для эффективного поиска контактов в телефонных книгах или каталогах.
6. Реализация ассоциативных массивов: TST может служить основой для реализации структур данных, подобных ассоциативным массивам, где ключами являются строки.

Основная суть тернарного дерева поиска заключается в том, что каждый узел хранит в себе один символ ключа, представляющего собой строку и может иметь до трёх дочерних узлов, а именно нижний (low), равный (equal) и верхний (high). Нижний и верхний узлы используются по тому же принципу, что левый и правый в бинарном дереве поиска, то есть в нижний дочерний узел попадает символ, который располагается ближе к началу алфавита, чем его родитель, а в верхний дочерний узел попадает символ, который находится дальше от начала алфавита, чем его родитель. Равный узел служит для того, чтобы хранить в себе те символы, которые являются продолжением хранящегося в дереве ключа. Помимо символа ключа и указателей на дочерние узлы в каждом узле хранится так же метка, хранящая в себе одно из двух возможных состояний узла, которая нужна для того, чтобы проверить, является ли символ, хранящийся в узле, завершающим символом ключа или нет.

# ОПИСАНИЕ ОСНОВНЫХ ОПЕРАЦИЙ

## 1. Вставка ключа в дерево

Для того, чтобы вставить в дерево новый ключ, требуется сначала найти то место, с которого следует начать вставку. Если дерево пустое, то новый ключ становится корнем дерева, а вот если в дереве уже были ключи, то начинается поиск узла node, с которого начнётся добавление нового ключа.

Изначально узлу node присваивается значение корня дерева. Во время поиска места под новый ключ, первый символ ключа сравнивается с теми символами, которые уже хранятся в дереве. Если символ нового ключа меньше (находится ближе к началу алфавита), чем тот ключ, который уже есть, то узел node заменяется на его нижний дочерний узел, а если символ больше (дальше от начала алфавита), чем тот, что уже добавлен, то узел node заменяется на верхний дочерний узел. При каждой замене узла node в переменную prev сохраняется его предыдущее состояние, а в переменную indicator сохраняется информация о том, на какой из дочерних узлов был заменён узел node для того, чтобы в дальнейшем можно было связать новые узлы с теми, что уже были добавлены. Если возникает такая ситуация, что первый символ ключа уже был добавлен в дерево, то узел node заменяется на свой равный дочерний узел, а исследуемый символ ключа заменяется на его следующий символ, после чего поиск продолжается дальше.

Если в ходе поиска будут найдены все символы ключа, то узел prev, в котором в таком случае хранится последний символ ключа, будет помечен как узел с завершающим символом вне зависимости от того, какая у него до этого была метка, и алгоритм завершит свою работу. Если до запуска алгоритма узел prev был помечен как узел без ключа, то теперь в дереве будет новый ключ, для которого уже были готовы все узлы, а если метка окончания ключа уже стояла, то повторная метка никак не повлияет на дерево.

После того, как место для новых узлов найдено, начинается добавление новых узлов в дерево. При этом первый новый узел может быть по отношению к его родителю как равным, так и нижним или верхним, а вот все следующие за ним узлы будут равными дочерними узлами.

Особенность алгоритма вставки ключей в тернарное дерево поиска заключается в том, что если в дереве уже хранится часть нового ключа, то те символы, которых не было в дереве, будут добавлены в нижнее или верхнее поддерево первого узла с несовпадающим символом, следующего за последним узлом с совпадающим символом. Таким образом, если узел является нижним или верхним по отношению к своему родителю, то символ, хранящийся в родительском узле, точно не является частью ключа, часть которого хранится в дочернем узле, но при этом другая часть ключа может хранится ближе к корню.

## 2. Удаление ключа из дерева

Удаление ключа является более сложной операцией, чем добавление. Во время удаления могут возникать различные неправильные состояния дерева, каждое из которых необходимо правильно обработать.

Удаление начинается с поиска узла, в котором хранится первый символ удаляемого ключа. Это действие происходит с помощью вспомогательной операции “start\_of\_key” (все вспомогательные операции будут описаны после операции удаления). Если такой узел был найден, то он сохраняется в переменную node, при этом в переменную prev сохраняется его родительский узел (если только первый символ ключа не хранится в корне, иначе в переменной prev ничего не хранится), а в переменную indicator сохраняется его отношение к родительскому узлу. Если такой узел не был найден, значит удаляемого ключа нет в дереве и алгоритм завершает свою работу.

Далее происходит поиск всех остальных символов ключа с использованием вспомогательной операции “end\_of\_key”. Помимо поиска данная операция сохраняет в массив key\_trace все узлы, которые лежат на пути от узла, хранящего в себе первый символ ключа до узла, хранящего в себе последний символ. Некоторые узлы на пути могут хранить в себе символы, которые не являются частью ключа из-за особенностей вставки ключей. Помимо этого, в массив indicators сохраняются отношения узлов к их родительским узлам. После выполнения “end\_of\_key” в переменную node сохраняется узел, хранящий в себе последний символ ключа (если он был найден). Если такой символ не был найден, значит соответствующий ключ отсутствует в дереве и алгоритм завершает свою работу.

После того, как все узлы, относящиеся к удаляемому ключу, были найдены, начинается последовательное удаление узлов, которое происходит, c помощью операции “free\_key”, начиная с последнего найденного узла, до тех пор, пока не встретится узел, принадлежащий более чем одному ключу или пока не будет удалён весь ключ, кроме первого символа ключа (узел с первым символом требуется удалять особенным образом). Также после выполнения операции “free\_key” в last\_node и last\_sym сохраняются индекс узла родителя последнего удалённого узла и индекс символа ключа, который ещё хранится в дереве.

Проверка осуществляется следующим образом: если какой-либо узел на пути от узла, хранящего в себе первый символ ключа, к узлу, хранящему в себе последний символ, помечен как узел, хранящий в себе последний символ какого-либо ключа или имеет дочерние узлы, кроме удаляемых, или является узлом, хранящим в себе первый символ удаляемого ключа или символ, которого нет в удаляемом ключе, то удаление прекращается.

После того, как были удалены все узлы, которые не нарушали условия, в переменную node сохраняется родитель последнего удалённого узла и начинается проверка нарушений свойств дерева, которые могли возникнуть в ходе удаления.

Первое условие проверяет, нет ли у node равного дочернего узла. Если такой узел есть, значит последний удалённый узел был нижним, либо верхним дочерним узлом node. В таком случае более никаких действий не требуется и алгоритм завершает свою работу.

Следующее условие проверяет, происходит ли удаление корня при удалении ключа из дерева. Если складывается такая ситуация, значит переменная prev ничего не хранит, при этом были удалены все символы ключа, кроме первого и первый символ ключа не является завершающим символом целого ключа.

Если условие выполняется и происходит удаление корня, то вызывается вспомогательная операция “tst\_restore\_prop\_any\_case”, которая перестраивает дерево для того, чтобы обозначить новый корень и сохранить бинарные свойства дерева. Подробности этой вспомогательной операции будут описаны позже. Далее алгоритм завершает свою работу.

Если не происходит удаление корня, то проверяется тот случай, при котором узел node помечен как конец ключа и не имеет дочерних узлов. В таком случае более никаких действий не требуется и возвращается корень дерева.

После этого происходит последняя проверка с одним условием, при котором node может быть узлом, хранящим первый символ ключа, либо любой другой символ. Если node хранит в себе первый символ ключа, то происходит восстановление свойств дерева вспомогательной операцией “tst\_restore\_prop”, которая принимает на вход индикатор, узел node и его родителя из массива key\_trace. В противном случае на вход подаются значения первого элемента indicators, переменой node и prev. После чего происходит завершение операции удаления.

## 2.1 Вспомогательные операции для удаления

### 2.1.1 start\_of\_key

Данная операция принимает на вход следующие данные: node – узел с которого начнётся поиск нужного узла, first\_sym – переменная, в которой хранится значение первого символа ключа, prev – переменная, в которую будет сохранён узел, предшествующий узлу с первым символом ключа и indicator – переменная, в которой будет хранится информация об отношении найденного узла node к узлу prev.

Внутри самой операции происходит поиск наподобие того, что происходит в бинарном дереве поиска, то есть, если символ меньше того, что хранится внутри узла, то узел переходит к нижнему дочернему узлу, а если символ больше, то к верхнему дочернему. Если же нужный дочерний узел отсутствует, значит в дереве отсутствует искомый первый символ ключа и поиск завершается с возвратом NULL. А если первый символ равен символу в каком либо из узлов во время поиска, то цикл поиска завершается и возвращается указатель на найденный узел.

### 2.1.2 end\_of\_key

Данная операция нужна для того, чтобы найти все узлы, хранящие в себе символы ключа. Функция принимает на вход переменную node, хранящую в себе узел с первым символом ключа, массив key\_trace, в котором будут хранится указатели на все узлы на пути от узла с первым символом до узла с последним символом, включая крайние узлы, массив indicators, который хранит информацию об отношении узлов из массива key\_trace к их родительским узлам и сам ключ key.

В самом начале функции переменной indicator присваивается значение первого элемента массива indicators, в котором к началу выполнения функции уже должна хранится информация об отношении узла node к его родительскому узлу. Также объявляется переменная key\_is\_part, которая понадобится для проверки принадлежности символа, хранящегося в каком-либо узле, к удаляемому ключу. Помимо этого, происходит объявление переменной n\_symbol, которая используется для обращения к символам ключа по индексу.

Далее начинается цикл для переменной i от нуля до значения внешней переменной max\_deep, которая используется при объявлении массивов key\_trace и indicators и задаёт их размер. Предполагается, что такого размера массивов будет достаточно. В начале цикла в массив key\_trace сохраняется узел node и его индикатор. Далее происходит проверка принадлежности символа ключа с индексом n\_symbol к узлу node. Если символ находится в узле, то происходит переход к следующему символу, то есть значение переменной n\_symbol увеличивается на единицу. А если символ в узле не соответствует символу ключа, то происходит поиск наподобие того, что в функции start\_of\_key, но при этом все узлы, которые появляются на протяжении поиска, сохраняются в массив key\_trace. Если символ не найден в дереве, то функция возвращает NULL, а если найден, то n\_symbol увеличивается на единицу и поиск завершается.

После поиска происходит две проверки. Сначала проверяется, есть ли ещё символы в ключе, и если их нет и при этом узел node помечен как завершающий, то функция заканчивает свою работу и возвращает узел node. Если же узел node не помечен как завершающий, значит ключ не найден, и функция возвращает NULL. А если в ключе ещё остались символы, то проверяется наличие равного дочернего узла node и переход к нему, если он есть, либо завершение с NULL, если его нет. После цикл для i переходит к следующей итерации. Если цикл дошёл до конца и завершил свою работу, значит в массиве key\_trace закончилось место, чего не должно произойти. В таком случае функция вернёт NULL.

### 2.1.3 free\_key

Следующая операция удаляет все узлы, которые получится, из тех, что были найдены в ходе работы функции end\_of\_key. Функция принимает на вход node (указатель не узел, хранящий в себе последний символ ключа, массив узлов key\_trace, массив индикаторов indicators, переменные last\_node, которая хранит в себе индекс последнего узла в массиве key\_trace, last\_sym, которая хранит индекс последнего символа в ключе и сам ключ key.

В самом начале работы функции узел node помечается как узел, который не хранит завершающий символ, после чего начинается цикл с множеством условий. Цикл выполняется пока у node нет дочерних узлов, узел node не помечен как узел с последним символом ключа, узел node не является первым символом ключа и узел node хранит в себе тот символ, который является частью удаляемого ключа.

В цикле происходит освобождение памяти, выделенной под узел node, после чего значение переменных last\_node и last\_sym уменьшается на единицу и узлу node присваивается адрес его родителя. Далее соответствующему полю узла node при помощи массива indicators присваивается значение NULL, чтобы удалить узел из дерева окончательно. После завершения цикла происходит возвращение узла node.

### 2.1.4 tst\_restore\_prop\_any\_case

Эта функция восстанавливает свойства дерева в тех случая, когда удаляемый узел их нарушает. Функция принимает на вход переменную node, которая хранит адрес узла, который нужно удалить, и переменную replacement, которая хранит в себе адрес того места в дереве, в котором требуется перестроить дерево после удаления узла node.

Функция начинается с проверки того, какие дочерние узлы имеются у узла node. Если у него есть верхний дочерний узел, то он встаёт на место node в переменную replacement, после этого происходит проверка наличия нижнего дочернего узла node. Если такого узла нет, то дерево перестроено и функция завершает свою работу, если такой узел есть, то его необходимо сделать нижним дочерним узлом меньшего узла в верхнем поддерево node. Для этого нижний дочерний узел node сохраняется в переменную lo\_node, после чего происходит поиск наименьшего узла в верхнем поддереве node. Поиск происходит следующим образом: в переменную hi\_node сохраняется указатель на верхний дочерний узел node, узел hi\_node заменяется на нижний дочерний узел до тех пор, пока такой имеется. После того, как был найден отсутствующий нижний дочерний узел, им становится тот узел, который был сохранён в переменную lo\_node. После чего дерево перестроено, и функция успешно завершает свою работу.

Если же у node нет верхнего дочернего узла, то происходит проверка на наличие нижнего дочернего узла. Если такой есть, то он просто встаёт на место узла node в переменной replacement и функция завершает свою работу.

Если у node не было дочерних узлов, то переменной replacement просто присваивается NULL.

### 2.1.5 tst\_restore\_prop

Данная функция является расширением для предыдущей. Она принимает на вход indicator, node и prev (родительский узел node). После чего она вызывает функцию tst\_restore\_prop\_any\_case, используя в качестве второго аргумента один из дочерних узлов prev, в зависимости от того, что хранится в переменной indicator.

## 3. Удаление всего дерева

Для удаления дерева применяется рекурсивная функция, которая принимает на вход корень дерева и вызывает саму себя для его дочерних узлов, если такие имеются. В конце функции происходит освобождение памяти, выделенной для каждого узла.

## 4. Вывод списка слов

### 4.1 tst\_print\_all\_words

Для вывода всех слов, которые хранятся в дереве применяется функция tst\_print\_all\_words, которая принимает на вход корень дерева.

Внутри функции создаётся строка размером max\_deep для буферизации слова, которое нужно будет вывести. После чего в строку word записывается первый символ, который лежит в корне дерева, и если корень дерева является узлом, помеченным, как завершающий, то word тут же выводится.

Далее, если у корня есть равный дочерний узел, то вызывается функция tst\_print\_words\_from\_node, которая выводит все завершения префикса. В качестве префикса этой функции на вход поступает единственный символ, который хранится в текущем узле.

После этого, если у корня есть нижний или верхний дочерний узлы, то функция tst\_print\_all\_words рекурсивно вызывается для этих узлов и процесс повторяется до тех пор, пока не закончатся все узлы, хранящие в себе первые символы слов.

### 4.2 tst\_print\_words\_from\_node

На вход функция принимает префикс word, узел node, который является дочерним узлом по отношению к тому, который хранит в себе последний символ префикса и переменную indicator, которая хранит в себе отношение узла node к его родительскому узлу.

Данная функция является рекурсивной, поэтому в начале функции создаётся копия префикса в строку word\_copy. После этого происходит проверка индикатора. Если node является верхним или нижним дочерним узлом своего родителя, значит последний символ word не является частью ключа. Это обусловлено особенностью добавления ключей, которая описывалась выше.

После этого к word\_copy в конец дописывается символ, хранящийся в узле node и происходит проверка на наличие метки завершения ключа у узла node. Если она есть, то word\_copy выводится.

Далее функция вызывает саму себя для всех дочерних узлов node, если они есть. При этом в качестве префикса в качестве аргумента используется word\_copy. Также используется индикатор, соответствующий дочернему узлу.

## 5. Поиск

Так как дерево не хранит в себе ничего, кроме ключей, а ключи при этом хранятся в разных узлах, было принято решение утроить поиск так, чтобы по итогу возвращалось значение true в случае, если ключ найден и значение false, если искомого ключа нет в дереве.

### 5.1 tst\_lookup

Данная функция занимается поиском ключа и принимает на вход указатель на корень дерева tree и ключ key.

Функция использует вспомогательную функцию tst\_find\_sym для того, чтобы правильно искать символы в дереве. Сначала при помощи этой функции происходит поиск первого символа ключа в дереве. После этого происходит переход к следующему символу и начинается цикл до конца ключа, если первый символ был найден в дереве.

Цикл начинается с проверки наличия равного дочернего узла у tree, где к моменту начала цикла сохранён узел с первым символом. Если такой дочерний узел есть, то происходит проверка, совпадает ли его символ с символом ключа. Если это так, то в tree записывается равный дочерний узел и происходит переход к следующему символу ключа, а цикл переходит к следующей итерации.

Если символ, который хранится в равном дочернем узле не совпадает с символом ключа, то в tree записывается адрес равного дочернего узла и цикл продолжает свою работу поиском символа ключа в дереве. Если символ найден, то произойдёт переход к следующему символу и новой итерации. Если символ не найден, то вернётся false. То же самое значение вернётся, если у tree отсутствует равный дочерний узел в начале цикла.

Если цикл завершил свою работу, значит для каждого символа ключа был найден соответствующий узел. Тогда функция возвращает метку последнего узла.

### 5.2 tst\_find\_sym

Функция принимает на вход указатель на узел node и символ ключа symbol. Она применяется в тех случаях, когда символ не найден в узле node и необходимо проверить его наличие в нижнем и верхнем поддереве узла node. Внутри функции происходит обычный поиск, подобие которого уже описывалось выше.

## 6. Префиксный поиск

Помимо обычного поиска, в рамках данной курсовой работы был реализован так же и префиксный поиск. Префиксный поиск позволяет найти все завершения n первых символов ключа (n может равняться количеству символов в ключе).

### tst\_prefix\_search

Данная функция принимает на вход указатель на корень дерева и префикс.

Первым делом создаётся копия префикса, которая понадобится в дальнейшем. После этого происходит поиск первого символа префикса в дереве. Далее происходит переход к следующему символу префикса и поиск остальных его символов в цикле. Данный поиск осуществляется так же, как и в функции tst\_lookup. Если найдены все символы префикса, то происходит вывод префикса, если узел с последним символом префикса помечен, как завершающий, после чего вызывается вспомогательная функция tst\_print\_words\_from\_node, которая выводит все завершения этого префикса, которые есть в дереве.

# АНАЛИЗ ЭФФЕКТИВНОСТИ АЛГОРИТМОВ