МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

ИСПОЛЬЗОВАНИЕ ЛОКАЛЬНО ЧУВСТВИТЕЛЬНОЙ ХЕШ-ФУНКЦИИ ДЛЯ СРАВНЕНИЯ И КЛАСТЕРИЗАЦИИ ПАКЕТОВ ПРОТОКОЛОВ СЕТЕВОГО УРОВНЯ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Гаджимурадова Руслана Магомедовича

Научный руководитель:
Доцент кафедры ИиП _____ Е. В. Кудрина

подпись, дата

Зав. кафедрой:
к.ф.-м.н., доцент _____ М.В. Огнева

подпись, дата

Саратов 2022

ВВЕДЕНИЕ

Актуальность темы. Сравнение различных объектов и определение их близости друг к другу – одни из тех задач, которые крайне актуальны в нынешнее время. Различные сервисы просмотра кино и прослушивания музыки сталкиваются с необходимостью рекомендации пользователям произведений, по вкусам и предпочтениям. Интернет-магазины стремятся предложить своим покупателям товары, имеющие больший потенциал к приобретению, на основании предыдущих покупок. Все эти задачи успешно решаются при помощи локально чувствительного хеширования.

Однако данный механизм позволяет при некоторой адаптации работать и с менее тривиальными объектами, например, такими как пакеты различных протоколов. Данная особенность может использоваться в компаниях, разрабатывающих оборудование, коммуницирующее по сети. Процесс взаимодействия такого оборудования логируется для возможности дальнейшего анализа исключительных ситуаций и сбоев. Программное обеспечение различных производителей позволяют, помимо ведения логов, собирать статистику, производить анализ и предупреждать об ошибках. Однако, такое оборудование не дает возможности работать с пакетами внутренних протоколов компаний. Можно предположить, что существует коммерческое программное обеспечение позволяющее производить более глубокую конфигурацию оборудования и последующий анализ. Тем не менее, маловероятно, что производители закладывают возможность конфигурации работы с нестандартными протоколами.

Локально чувствительное хеширование позволяет решить описанную проблему, так как оно не привязано к конкретным протоколам или их версиям, а является универсальным механизмом. Еще большую значимость данному подходу придает тот факт, что новые протоколы или версии старых могут разрабатываться каждый раз при проектировании и выпуске компаниями новых изделий, коммуницирующих по сети, а следовательно, требуется универсальное решение. Также стоит отметить, что не все

устройства могут сообщаться посредством витой пары и соответствующими логическими уровнями связи. На производстве, зачастую, используются иные интерфейсы, такие как: RS-232, RS-485, CAN и другие. Локально чувствительное хеширование, реализованное в программном обеспечении верхнего уровня, может быть использовано для работы с пакетами и этих стандартов.

Еще одним преимуществом локально чувствительного хеширования пакетов протоколов является возможность уменьшения занимаемого логами объема памяти путем архивации полученных пакетов. При этом в память дополнительно записываются подсчитанные значения хеш-функции каждого пакета, по которым, при необходимости, производится изначальный поиск или кластеризация.

Цель бакалаврской работы – исследование возможности применения локально чувствительного хеширования для сравнения и кластеризации пакетов протоколов сетевого уровня.

Поставленная цель определила **следующие задачи**:

1. Изучить теоретические основы локально чувствительного хеширования;
2. Рассмотреть возможность применения локально чувствительного хеширования пакетов протоколов сетевого уровня для решения задач сравнения и кластеризации;
3. Представить обзор инструментальных средств, применяемый для хеширования пакетов протоколов сетевого уровня;
4. Реализовать локально чувствительное хеширование и применить его для решения задач сравнения и кластеризации пакетов протоколов сетевого уровня.

Методологические основы использования локально чувствительной хеш-функции для сравнения и кластеризации пакетов протоколов сетевого уровня представлены в работах Э. Таненбаума [1], Ю. Лесковеца [2].

Теоретическая значимость бакалаврской работы заключается в исследовании возможности применения локально чувствительного

хеширования для работы с пакетами протоколов сетевого уровня и выборе правильной метрики различия полученных хешей для решения задач сравнения и кластеризации.

Практическая значимость бакалаврской работы заключается в реализации локально чувствительной хеш-функции пакетов протоколов сетевого уровня и ее применении при решении задач сравнения и кластеризации.

Структура и объём работы. Бакалаврская работа состоит из введения, двух разделов, заключения, списка использованных источников и трех приложений. Общий объем работы – 72 страницы, из них 44 страницы – основное содержание, включая 1 рисунок и 20 таблиц. Список использованных источников информации содержит 23 наименования. Код реализованных алгоритмов представлен в качестве приложения.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Теоретическая часть» бакалаврской работы посвящен теоретическим основам локально чувствительного хеширования и возможности его применения для решения задач сравнения и кластеризации пакетов протоколов сетевого уровня. В данном разделе рассматриваются стандарт протоколов сетевого уровня, а также структуры пакетов таких протоколов как: ARP, IPv4, ICMP.

Рассматриваемый в данном разделе Address Resolution Protocol служит для определения MAC-адреса устройств по известному IP-адресу. Пакет протокола ARP имеет структуру, представленную в таблице 1.

Таблица 1 – Структура пакета протокола ARP

Биты	**Биты с 0 по 7**	**Биты с 8 по 15**	**Биты с 16 по 31**
0	HTYPE		PTYPE
32	HLEN	PLEN	OPER
64	SHA		
-	SPA		

Продолжение таблицы 1

Биты	Биты с 0 по 7	Биты с 8 по 15	Биты с 16 по 31
-	THA		
-	TPA		

Пакеты протокола IPv4 состоят из заголовка и части данных. Заголовок, в свою очередь, делится на фиксированную часть и опциональную. Фиксированная часть имеет длину 20 байт и соответствует структуре, представленной в таблице 2.

Таблица 2 – Структура пакета протокола IP

Октеты	Биты	0	32	64	96	128	160	…	448
0	0	Version	Identification	TTL	Source IP Address	Destination IP Address	Options		
	1								
	2								
	3								
	4	IHL							
	5								
	6								
	7								
	8								
1	9			Protocol					
	10								
	11								
	12								
	13								
	14	ECN							
	15								
2	16	Total Length	Flags	Header Checksum					
	17								
	18								
	19		Fragment Offset						
	20								
	21								
	22								
	23								
3	24								
	25								
	26								
	27								
	28								

Продолжение таблицы 2

Октеты	Биты	0	32	64	96	128	160	…	448
	29								
	30								
	31								

Internet Control Message Protocol служит для передачи сообщений об ошибках и различных событий, возникающих при передачи сообщений по сети. Структура пакета протокола ICMP представлена в таблице 3.

Таблица 3 – Структура пакета протокола ICMP

Октеты	Биты	0	32
0	0	Type	Rest of header
	1		
	2		
	3		
	4		
	5		
	6		
	7		
1	8	Code	
	9		
	10		
	11		
	12		
	13		
	14		
	15		
2	16	Checksum	
	17		
	18		
	19		
	20		
	21		
	22		
	23		
3	24		
	25		
	26		
	27		
	28		
	29		
	30		
	31		

В данном разделе дается определение локально чувствительной хеш-функции, как алгоритму понижения размерности данных с целью их дальнейшего сравнения путем помещения схожих объектов в один набор с высокой степенью вероятности. Однако, в отличие от изображений или музыкальных записей, пакеты протоколов, рассматриваемые в данной работе, не несут в себе излишней или потенциально ненужной информации.

При этом размеры пакетов теоретически могут достигать 65535 байт. В связи с перечисленными выше особенностями, требуются дополнительные действия для уменьшения размерности пакета при сохранении значимой информации. В данном разделе работы определен следующий алгоритм подсчета значения локально чувствительной хеш-функции:

1. Сжатие и полей и распределение по важности;
2. Применение операции побитового исключающего или;
3. Умножение результата на степень двойки;
4. Разбиение и суммирование.

Для сравнения полученных сигнатур в работе были определены используемые метрики: коэффициент Жаккара, Евклидова метрика, собственная метрика

Мера Жаккара задается отношением мощности пересечениях двух множеств к мощности объединения этих множеств:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Евклидова пространства данная метрика задается формулой:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

Собственная метрика задается следующим равенством:

$$m(P,Q) = \frac{1}{n} \sum_{p_i \in P, q_i \in Q} \frac{|p_i - q_i|}{M}$$

где P и Q – сигнатуры двух пакетов, полученные применением локально чувствительной хеш-функции, n – число элементов каждой из сигнатур, $M = 2^{16} - 1$.

Таким образом, в первом разделе определена структура пакетов, алгоритм подсчета их локально чувствительной хеш-функции и метрики сравнения.

Второй раздел «Практическая часть» бакалаврской работы посвящен реализации представленного в первом разделе алгоритма нахождения

значения локально чувствительной хеш-функции для пакетов протоколов сетевого уровня.

На примере пакета протокола ARP, значение полей которого, а также их важность и требуемая точность представлены в таблице 4, показано полученное значение хеша.

Таблица 4 – Пакет ARP из предыдущего примера, но с другим набором полей и важность, а также полученная сигнатура

Поле	**Значение**	**Важность**	**Точность**
HTYPE	0x01	Низкая	Младшие Биты
PTYPE	0x08 0x00	Низкая	Младшие Биты
HLEN	0x06	Низкая	Младшие Биты
PLEN	0x04	Низкая	Младшие Биты
OPER	0x02	Высокая	Младшие Биты
SHA	0xED 0x10 0x38 0x9E 0xED 0xA4	Низкая	Старшие Биты
SPA	0x10 0xFD 0xA8 0xC0	Средняя	Старшие Биты
THA	0x82 0xC1 0x2F 0x98 0x75 0x94	Низкая	Старшие Биты
TPA	0xA5 0xE3 0xA8 0xC0	Средняя	Старшие Биты
Сигнатура			
512	0	101	0

В данном разделе бакалаврской работы представлена реализация решение задачи поиска ближайшего соседа. Результат работы решения данной задачи с помощью собственной метрики показан на примере представленном в таблице 5.

Таблица 5 – Результат работы метода поиска ближайшего соседа для пакета протокола IPv4

IHL	**DSCP**	**ECN**	**Total Length**	**Flags**	**Offset**	**TTL**	**Source IP**	**Destination IP**	**Различия**
20	0	0	116	0	0	255	192.168.13.6	192.168.13.1	0
20	0	0	116	0	0	255	192.168.13.6	192.168.13.1	0
20	0	0	116	0	0	255	192.168.13.6	192.168.13.1	0
20	0	0	114	0	0	255	192.168.13.6	192.168.13.1	0,001907

Продолжение таблицы 5

IHL	DSCP	ECN	Total Length	Flags	Offset	TTL	Source IP	Destination IP	Различия
20	0	0	114	0	0	255	192.168.13.6	192.168.13.1	0,001907
20	0	0	112	0	0	255	192.168.13.6	192.168.13.1	0,003815

В данном разделе представлена реализация метода решения задачи кластеризации, результаты работы которого отображены на рисунке 1. Из представленного рисунка видно, как 1900 пакетов протокола IPv4 были распределены на 20 кластеров. В кластеры 19 и 20 попало наибольшее число элементов. Это связано с тем, что к данным кластерам были отнесены пакеты IPv4 без вложений, которых оказалось больше, чем пакетов IPv4 с вложенным пакетом ICMP.

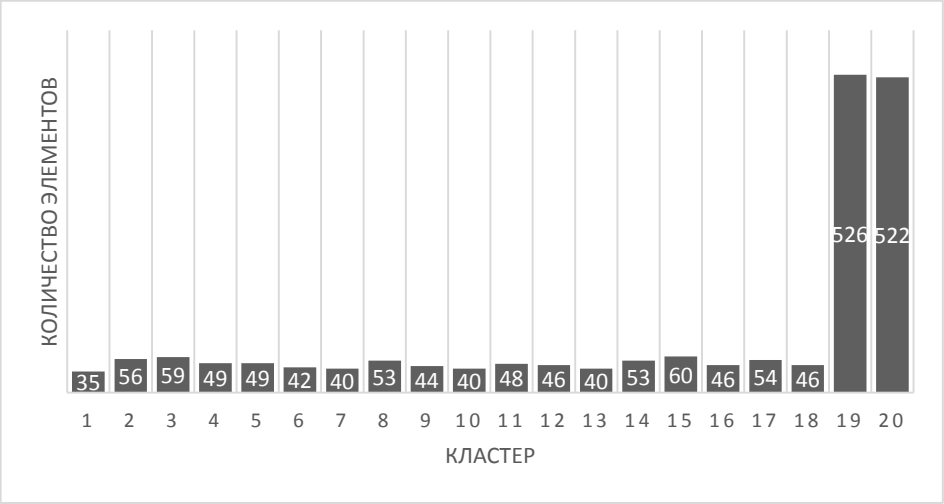

Рисунок 1 – Диаграмма полученных кластеров и числа их элементов

Таким образом, во втором, практическом, разделе бакалаврской работы представлена программная реализация описанной в первом разделе локально чувствительной хеш-функции, метрик сравнения полученных сигнатур и алгоритмов поиска ближайшего соседа и кластеризации. Также в данном разделе на конкретных примерах отображены результаты работы реализованных методов, визуализированные в таблицах и диаграмме.

ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы все поставленные задачи были выполнены, что позволило исследовать возможность применения локально чувствительного хеширования пакетов протоколов сетевого уровня для решения практических задач.

Результаты данной работы доказывают, что локально чувствительное хеширование может быть использовано при работе со столько нетривиальными и структурированными объектами, как пакеты протоколов. Стоит отметить, что специфика, реализованной в данной работе хеш-функции такова, что она может использоваться для работы с пакетами протоколов не только сетевого уровня, но и с пакетами любых других протоколов. Возможность указать важность каждого поля позволяет достаточно гибко настроить влияние значения тех или иных полей как на итоговую сигнатуру, так и на параметры, на основании которых производится кластеризация.

Основные источники информации:

1. Таненбаум Э. Computer networks / Э. Таненбаум, Д. Уэзеролл. – М.: Pearson Education, 2021. - 435 с.
2. Лесковец Ю. Mining of Massive Datasets / Ю. Лесковец, Раджараман А., Ульман Д. – М.: Cambridge University Press, 2020. - 565 с.
3. ARP Packet Format [Электронный ресурс]: статья – URL: https://www.javatpoint.com/arp-packet-format (дата обращения 31.05.2022).
4. IPv4 vs IPv6: What is the Difference Between IPv4 and IPv6? [Электронный ресурс]: статья. – URL: https://clck.ru/prrb4 (дата обращения 10.05.2022).
5. Internet Protocol Version 4 (IPv4) Parameters [Электронный ресурс]: статья. – URL: https://www.iana.org/assignments/ip-parameters/ip-parameters.xhtml (дата обращения 10.05.2022).
6. Оценка качества в задаче кластеризации [Электронный ресурс]: статья – URL: https://cutt.ly/5JkHgZ0 (дата обращения 25.05.2022).