УДК 621.391

Н. В. КОЖЕМЯКИНА, Н. Н. ПОНОМАРЕНКО

Национальный аэрокосмический университет им. Н. Е. Жуковского «ХАИ», Украина

МЕТОД СЖАТИЯ ДАННЫХ МОНИТОРИНГА ТРАФИКА ДЛЯ СРЕДСТВ ТЕЛЕКОММУНИКАЦИЙ

Рассмотрена задача сжатия данных, содержащих информацию об основных параметрах сетевого трафика. Сформированы двенадцать тестовых наборов с различными видами сетевого трафика для известных утилит Wireshark, Colasoft Capsa и CommView. Показано, что при сжатии основной объем памяти в сжатых данных приходится на отметки времени. Предложен метод сжатия отметок времени, состоящий из вычисления дельт, преобразования Барроуза-Уилера (BWT), кодирование расстояний (DC) и рекурсивного группового кодирования (РГК) на завершающей стадии. Показано, что использование РГК на завершающей стадии обеспечивает более эффективное кодирование, чем известные методы. Показано, что предложенный метод сжатия обеспечивает коэффициент сжатия отметок времени более чем в 2 раза выше, чем WinRar.

***Ключевые слова**: сжатие данных, методы анализа и мониторинга трафика, преобразование Барроуза-Уилера, кодирование расстояний.*

Введение

С расширением предоставляемых услуг в сфере телекоммуникаций и одновременным улучшением их качества, растет объем данных, передаваемых в современных сетях связи, и количество пользователей этими данными. И за этим стремительным ростом едва успевают современные сети связи. Речь идет не только об увеличении загруженности каналов связи с ростом, например, мультимедийного трафика, но и появлении новых приложений и протоколов, позволяющих оперировать все большим количеством информации, а также об увеличении количества гаджетов, обменивающихся m2m трафиком, что приводит к усложнению архитектуры сетей и увеличению объемов передаваемых данных.

Наряду с увеличением объемов передаваемой полезной информации, возрастает и количество передаваемых служебных данных. Все это приводит к увеличению количества передаваемых пакетов. Это не только загружает каналы связи, но и усложняет работу по проектированию и отладке сетей, затрудняет работу их администраторов. Появляется необходимость в более тщательном анализе передаваемого трафика, для чего необходимо осуществлять его мониторинг и сохранять его данные [1]. При этом объемы сохраняемых данных мониторинга могут достигать очень больших размеров. Файл сжатых данных, сохраняемых для одного обычного узла сети в течение часа, может составлять от 40 MB для протокола прикладного уровня HTTP, предоставляющего доступ к сетевым службам, до 6 GB для трафика протокола сетевого уровня BitTorrent. При этом количество таких узлов, на которых осуществляется мониторинг трафика, может измеряться сотнями и тысячами. Все это делает задачу разработки и совершенствования методов сжатия данных мониторинга трафика чрезвычайно актуальной [2].

Современные архиваторы имеют хорошие показатели сжатия, однако при анализе коэффициентов сжатия отдельных видов трафика видно, что эффективность сжатия для них сильно отличается (в десятки раз). *В работе предложен метод предварительного преобразования таких данных с целью улучшения коэффициента их сжатия.* Для устранения статистической и словарной избыточности в таких данных на разных этапах сжатия могут использоваться как сложные составные методы сжатия (архиваторы WinZip, WinRar), так и различные преобразования (BWT [3, 4], DC [4]) и элементарные базовые методы сжатия, такие как арифметическое кодирование (АК), кодирование Хаффмана (КХ). В работе анализируется возможность и эффективность использования в составе таких методов сжатия нового эффективного метода РГК [5]. РГК показывает себя эффективным при сжатии однородных по составу данных больших алфавитов [6], которые характерны для сжатия, в частности, мультимедийного трафика [7].

В подразделе 1 данной работы описывается процесс формирования тестовых наборов с помощью известной утилиты мониторинга трафика Wireshark [8], Colasoft Capsa и CommView. В подразделе 2 анализируется сжимаемость различных составляющих данных мониторинга трафика и предлагается новый метод сжатия отметок времени. В подразделе 3 осуществляется сравнительный анализ различных методов сжатия, применяемых на последнем шаге предложенного метода.

1. Формирование тестовых наборов

Наборы тестовых данных для оценки эффективности сжатия предлагаемого метода и других архиваторов можно сформировать несколькими способами. Можно создать «искусственный» трафик, воспользовавшись программами генерации сетевого трафика. Однако данный трафик будет значительно отличаться от передаваемого в реальных сетях связи. Вторым способом формирования тестовых наборов является использование сетевых анализаторов (так называемых снифферов), которые работают непосредственно с трафиком сетевых интерфейсов. Основной функцией сниффера является перехват пакетов, их декодирование, отображение содержимого пакетов для дальнейшего анализа и обработки сетевого трафика.

При обзоре современных сетевых анализаторов, (программных снифферов), выбор был остановлен на таких программах, как Wireshark, Colasoft Capsa и CommView. Так, например, анализатор Wireshark обладает удобным графическим интерфейсом, поддерживает работу с более чем 1000 сетевыми протоколами, имеет различные настраиваемые фильтры, совместим с другими программами-аналогами и свободен в распространении. Для возможности анализа и работы с потоками данных снифферы предоставляют информацию об источнике и получателе пакета, его длине, номере используемого порта, типе протокола, длительности, времени отправки и получения пакетов и их содержимом.

Для возможности оценки эффективности разрабатываемого метода сжатия и проведения сравнительного анализа при помощи сетевого анализатора были сформированы 12 наборов тестовых данных. Первый набор состоит из данных протокола прикладного уровня HTTP, предоставляющего доступ к сетевым службам. Второй набор был сформирован в результате перехвата информации с портов, передающих пакеты по протоколу сетевого уровня BitTorrent (обмен файлами в сети). Третий и четвертый наборы состоят из пакетов протоколов TCP (установка соединения с гарантированной доставкой) и UDP (пересылка дейтаграмм без исправления ошибок) транспортного уровня, которые являются основными в IP сетях, и отличаются по структуре заголовка пакета и максимальным количеством полезной информации, передаваемой одним пакетом. Первые четыре набора содержат по сто тысяч пакетов. Пятый и шестой набор, в отличие от вышеперечисленных, состоят из одного миллиона пакетов и представляют собой трафик протокола BitTorrent. Описанные выше шесть тестовых наборов были сформированы при помощи анализатора Wireshark. Следующие три набора были получены снифером CommView и представляют собой набор пакетов, полученных при перехвате трафика без использования каких-либо фильтров и ограничений (седьмой набор) и наборов, состоящих из пакетов протоколов TCP (восьмой набор) и UDP (девятый набор). Подобно предыдущим трем наборам были сформированы десятый, одиннадцатый и двенадцатый наборы, только для их создания использовался сниффер Colasoft Capsa.

Данные из полученных наборов были сгруппированы по типу информации и помещены в отдельные файлы. Таким образом, из каждого набора данных были получены 4 отдельных файла с информацией об отправителе пакета, его получателе, длине пакета и времени. Далее эти файлы были сжаты рядом широко используемых на данный момент архиваторов. В таблицах 1 приведены коэффициенты сжатия архиваторами WinRAR и WinZIP файлов, образованных из наборов данных, полученных при помощи Wireshark, CommView, Colasoft.

Таблица 1

Коэффициенты сжатия тестовых файлов

Наборы	Коэффициент сжатия							
	WinRAR				WinZIP			
	Time	Source	Destination	Length	Time	Source	Destination	Length
1	5,28	38,90	35,97	11,75	4,70	35,95	33,51	11,55
2	4,58	26,19	25,97	10,98	4,24	25,10	24,43	10,66
4	5,17	54,92	54,02	8,29	4,55	53,64	53,04	7,92
7	4,49	29,39	26,53	8,16	4,16	24,87	23,06	7,74
8	4,58	27,56	24,86	8,04	4,25	23,61	21,68	7,58
9	4,70	68,89	68,37	8,95	4,34	68,37	67,89	8,45
10	3,42	25,30	22,56	7,56	3,16	22,88	21,05	6,93
11	3,76	28,13	24,75	7,56	3,50	25,49	22,60	6,93
12	3,37	34,13	37,39	7,21	3,17	32,02	35,27	6,71

Как видно из таблицы 1, архиваторы хорошо справились со сжатием информации об отправителе и получателе, хуже с длиной пакета. Однако коэффициенты сжатия данных о времени получения и отправки пакетов оказались сжатыми хуже всего. Таким образом, самой сложной для сжатия оказывается именно информация об отметках времени.

На рис. 1 показана гистограмма, показывающая, какой объем в сжатых данных соответствует различным типам данных. Как видно, данные об отметках времени, не только сжимаются хуже всего, но и составляют почти 70% от сжатых данных.

Соответственно, из всех данных, содержащихся в полученных пакетах, наибольший интерес для

разработки нового метода сжатия представляет информация о метках времени получения и отправки пакетов. Она представляет собой массив данных, в котором каждый следующий элемент по значению больше предыдущего, т.е. значения элементов постоянно увеличиваются.

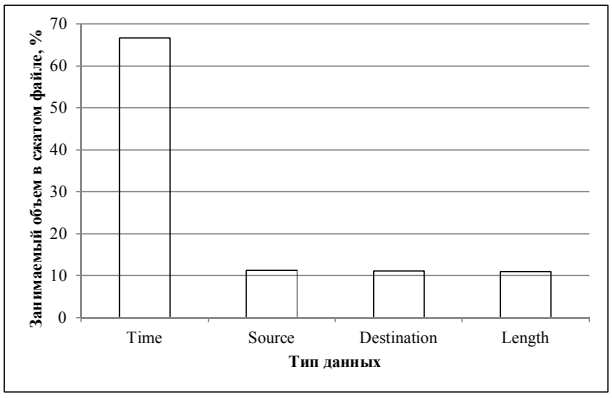

Рис. 1. Средний вклад различных типов данных в сжатых WinRar данных

Именно для этого типа данных разработка нового более эффективного метода сжатия способна привести к наибольшему увеличению общего коэффициента сжатия данных мониторинга трафика.

2. Предлагаемый метод сжатия

На рис. 2 приведена предлагаемая схема кодирования и сжатия данных.

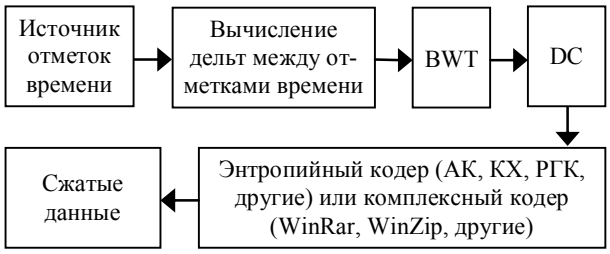

Рис. 2. Блок-схема предлагаемого метода сжатия

На первом шаге вычисляются дельты между соседними отсчетами времени. Как показывает практика, в подавляющем большинстве ситуаций эти дельты являются небольшими по сравнению с исходными положительными числами (очень редко и видимо из-за сбоев в WireShark встречаются отрицательные дельты). На втором шаге эти дельты подвергаются BWT, которое позволяет учесть словарные подобия в данных. На третьем шаге применяется DC, которое обычно применяется в связке с BWT после него. На последнем этапе сжатия необходимо устранить статистическую избыточность в данных. Для этого можно применить один из элементарных методов сжатия, такой как АК или КХ. Вместо АК или КХ можно применить разработанный недавно метод РГК. Этот метод, показывая в худшем случае эффективность сжатия, близкую к АК [5], для сжатия однородных данных больших алфавитов показывает существенно более высокие коэффициенты сжатия [7] при меньшей вычислительной сложности процесса сжатия. К таким данным, в частности, относятся квантованные блоки ДКП при сжатии изображений с потерями. В этом случае размеры символов алфавита могут достигать 128 байт [9, 10], а в ряде случаев 2048 байт [11, 12] или даже больше. Для данной задачи после DC данные будут являться двухбайтными, и необходим метод устранения статистической избыточности, который был бы эффективным для их кодирования и в то же время не требовательным к занимаемой памяти и быстрым. Таким методом и является РГК.

Кроме вышеперечисленных методов на последнем шаге можно попытаться применить и сложный метод сжатия, например, WinRar или WinZip. Таким образом, можно будет сравнить их эффективность для решения этой задачи с эффективностью АК, КХ и РГК.

3. Анализ данных

В таблице 2 приведены размеры файлов с метками времени, полученные из исходных тестовых наборов, а также размеры данных файлов после применения к ним преобразований всех шагов сжатия, кроме последнего (после DC).

Таблица 2

Размеры файлов до и после преобразований

Наборы данных	Размер исходного файла, байт	Размер преобразованного файла, байт
1	1600136	528823
2	1399924	509188
3	1677926	590940
4	1605127	591196
5	14815481	5298854
6	14799430	5189940
7	712778	280121
8	731684	281947
9	749015	252055
10	513299	340969
11	391233	223727
12	464084	283199

Как видно из данных таблицы 2, уже после DC размер файла несколько уменьшается, хотя в нем остается большая статистическая избыточность.

В таблице 3 приведены коэффициенты сжатия преобразованных файлов меток предложенным ме-

тодом сжатия с различными методами сжатия на завершающем этапе. Коэффициенты сжатия приведены по отношению к исходным несжатым файлам и могут быть сравнены с данными таблицы 1. Для облегчения анализа приведена и теоретическая достижимая степень сжатия по теореме Шеннона при кодировании данных, интерпретируемых как символы однобайтного алфавита.

Таблица 3

Коэффициенты сжатия файлов с метками времени

Тестовые файлы	Метод кодирования					
	Шеннон	РГК	АК	КХ	WinRAR	WinZIP
1	9,44	11,32	9,33	9,13	8,43	8,26
2	8,93	10,80	8,81	8,63	8,12	7,87
3	8,89	10,68	8,79	8,60	8,06	7,80
4	8,48	10,18	8,38	8,20	7,62	7,44
5	9,16	11,22	9,15	8,85	8,49	8,15
6	9,35	11,45	9,34	9,03	8,63	8,32
7	7,93	9,42	7,75	7,67	7,30	6,94
8	8,09	9,61	7,91	7,82	7,45	7,09
9	9,26	10,94	9,02	8,96	8,15	8,08
10	4,75	5,70	4,66	4,42	4,17	4,18
11	5,51	6,54	5,35	5,33	4,96	4,82
12	5,08	6,05	4,96	4,92	4,56	4,45

Как видно из данных таблицы 3, наибольшую эффективность сжатия обеспечивает метод РГК. Обеспечиваемые им коэффициенты сжатия более чем в 2 раза превышают коэффициенты сжатия для WinRar из таблицы 1. При этом РГК за счет эффективного кодирования символов больших алфавитов выигрывает у АК в среднем в 1,22 раза в степени сжатия, и у КХ в 1,25 раза. WinRar и WinZip для этой задачи (устранение статистической избыточности на последнем этапе сжатия) показывают худшую эффективность, чем элементарные методы устранения статистической избыточности, показывая как меньшую степень сжатия, так и меньшее быстродействие.

Заключение

Таким образом, наилучшим из проанализированных методов для использования на последнем этапе предложенного метода сжатия является РГК. Предложенный метод сжатия за счет более эффективного кодирования отметок времени способен уменьшить объем сжатых данных мониторинга трафика в среднем на 35%.

Литература

1. A comparative analysis of web and Peer-to-Peer traffic [Text] / N. Basher, A. Mahanti, A. Mahanti, C. Williamson, M. Arlitt // International Conference on World Wide Web. – Beijing, China, 2008. – P. 287-296.

2. Marcelloni, F. An efficient lossless compression algorithm for tiny nodes of monitoring wireless sensor networks [Text] / F. Marcelloni, M. Vecchio // Computer Journal. – 2009. – Vol. 52, N. 8. – P. 969-987.

3. Burrows, M. A block sorting lossless data compression algorithm. Technical Report 124: Digital Equipment Corporation [Text] / M. Burrows, D. Wheeler. – Systems Research Center – 1994. – 24 p.

4. Adjeroh, D. The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching [Text] / D. Adjeroh, T. Bell, A. Mukherjee. – Springer Science & Business Media. – 2008. – 352 p.

5. Fast recursive coding based on grouping of Symbols [Text] / N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola // Telecommunications and Radio Engineering. – 2009. – Vol. 68, N 20. – P. 1857-1863.

6. Ryabko, B. Fast codes for large alphabets [Text] / B. Ryabko, J. Astola, K. Egiazarian // Communications in information and systems. – October, 2003. – Vol. 3, N 2. – P. 65-78.

7. Means and results of efficiency analysis for data compression methods applied to typical multimedia data [Text] / N. Kozhemiakina, N. Ponomarenko, V. Lukin, K. Egiazarian, J. AStola // IEEE First International Scientific-Practical Conference Problems. – 2014. – P. 12-14.

8. Network forensics analysis using Wireshark [Text] / V. Ndatinya, Z. Xiao, V.R. Manepalli, K Meng, Y. Xiao // International Journal of Security and Networks. – 2015. – Vol. 10, N 2. – P. 91-106.

9. DCT based high quality image compression [Text] / N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola // Scandinavian Conference on Image Analysis. – Copenhagen, Denmark, 2005. – P. 1177-1185.

10. Additional lossless compression of JPEG images [Text] / N. Ponomarenko, K. Egiazarian, V. Lukin, J. Astola // Proceedings of 4th Symposium on Image and Signal Processing and Analysis. – Zagreb, Croatia, 2005. – P. 117-120.

11. Ponomarenko, N. ADCTC: A new high quality DCT based coder for lossy image compression [Electronic resource] / N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola. – 80 Min / 700 MB. CD ROM Proceedings of LNLA. – Switzerland, August – 2008. – 6 p. – 1 electronic optical disc (CD-ROM).

12. Bazhyna, A. V. Efficient bit-planes based method for compression of 3D-DCT coefficients [Text] / A. V. Bazhyna, K. O. Egiazarian, N. N. Ponomarenko // Proceedings of Picture Coding Symposium, Lisboa, Portugal, 7-9 November, 2007. – P. 4.

Reference

1. Basher, N., Mahanti, A., Mahanti, A., Williamson, C., Arlitt, M. A comparative analysis of web and Peer-to-Peer traffic. *International Conference on World Wide Web*, Beijing, China, 2008, pp. 287-296.

2. Marcelloni, F., Vecchio, M. An efficient lossless compression algorithm for tiny nodes of monitoring wireless sensor networks. *Computer Journal*, 2009, vol. 52 no. 8, pp. 969-987.

3. Burrows, M., Wheeler, D. *A block sorting lossless data compression algorithm. Technical*

Report 124: Digital Equipment Corporation. Systems Research Center Publ., 1994. 24 p.

4. Adjeroh, D., Bell, T., Mukherjee, A. *The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching*. Springer Science & Business Media Publ., 2008. 352 p.

5. Ponomarenko, N., Lukin, V., Egiazarian, K., Astola, J. Fast recursive coding based on grouping of Symbols. *Telecommunications and Radio Engineering*, 2009, vol. 68, no. 20, pp. 1857-1863.

6. Ryabko, B., Astola, J., Egiazarian, K. Fast Codes for Large Alphabets. *Communications in Information and Systems*, 2003, vol.3, no. 2, pp.139-152.

7. Kozhemiakina, N., Ponomarenko, N., Lukin, V., Egiazarian, K., Astola, J. Means and results of efficiency analysis for data compression methods applied to typical multimedia data. *International Scientific-Practical Conference Problems of Infocommunications Science and Technology*, Kharkov, Ukraine, 2014, pp. 12-14.

8. Ndatinya, V., Xiao, Z., Manepalli, V., Meng, K, Xiao, Y. Network forensics analysis using Wireshark. *International Journal of Security and Networks*, 2015, vol. 10 no. 2, pp. 91-106.

9. Ponomarenko, N., Lukin, V., Egiazarian, K., Astola, J. DCT based high quality image compression. *Scandinavian Conference on Image Analysis*, Copenhagen, Denmark, 2005, pp. 1177-1185.

10. Ponomarenko, N., Egiazarian, K., Lukin, V., Astola, J. Additional lossless compression of JPEG images. *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, Zagreb, Croatia, 2005, pp. 117-120.

11. Ponomarenko, N., Lukin, V., Egiazarian, K., Astola, J. ADCT: a new high quality DCT based coder for lossy image compression. *Proceedings of International Workshop on Local and Non-Local Approximation in Image Processing (LNLA '08),* Lausanne, Switzerland, 2008, pp. 1-6. CD-ROM.

12. Bazhyna, A., Egiazarian, K., Ponomarenko, N. Efficient bit-planes based method for compression of 3D-DCT coefficients. *Proceedings of Picture Coding Symposium*, Lisboa, Portugal, 2007, 4 p.

Поступила в редакцию 10.02.2016, рассмотрена на редколлегии 18.02.2016

МЕТОД СТИСНЕННЯ ДАНИХ МОНІТОРИНГУ ТРАФІКУ ДЛЯ ЗАСОБІВ ТЕЛЕКОМУНІКАЦІЙ

Н. В. Кожемякіна, М. М. Пономаренко

Розглянуто задачу стиснення даних, що містять інформацію про основні параметри мережевого трафіку. Сформовано дванадцять тестових наборів з різними видами мережевого трафіку для відомих утиліт Wireshark, Colasoft Capsa і CommView. Показано, що при стисненні основний обсяг пам'яті в стислих даних доводиться на позначки часу. Запропоновано метод стиснення відміток часу, що складається з обчислення дельт, перетворення Барроуза-Уілера (BWT), кодування відстані (DC) і рекурсивного групового кодування (RGC) на завершальній стадії. Показано, що використання RGC на завершальній стадії забезпечує більш ефективне кодування, ніж відомі методи. Показано, що запропонований метод стиснення забезпечує коефіцієнт стиснення відміток часу більш ніж у 2 рази вище, ніж WinRar.

Ключові слова: стиснення даних, методи аналізу та моніторингу трафіку, перетворення Барроуза-Уілера, кодування відстаней.

METHOD OF DATA COMPRESSION FOR TRAFFIC MONITORING TOOLS OF COMMUNICATION

N. V. Kozhemiakina, N. N. Ponomarenko

In this paper a problem of compressing data containing information on basic parameters of network traffic is considered. Twelve test sets with different types of network traffic for known monitoring tools Wireshark, Colasoft Capsa and CommView are formed. It is shown that the main part of memory in compressed data relates to timestamps. A method for compressing timestamps that consists in delta calculation, Burrows-Wheeler transform (BWT), distance coding (DC) and recursive group coding (RGC) at the final stage is proposed. It is demonstrated that the use of RGC at the final stage provides more efficient coding compared to known methods. It is also shown that the proposed method of timestamps coding produces about twice larger compression ratio than WinRar.

Key words: data compression, traffic monitoring and analysis tools, BWT, distance coding.

Кожемякина Надежда Владимировна – аспирант каф. приема, передачи и обработки сигналов, Национальный аэрокосмический университет им. Н. Е. Жуковского «Харьковский авиационный институт», Харьков, Украина, e-mail: nadejda_kozickaya@mail.ru.

Пономаренко Николай Николаевич - д-р техн. наук, доцент, профессор каф. приема, передачи и обработки сигналов, Национальный аэрокосмический университет им. Н. Е. Жуковского «Харьковский авиационный институт», Харьков, Украина, e-mail: nikolay@ponomarenko.info.

Kozhemiakina Nadejda Vladimirovna - PhD Student, Department of Transmitters, Receivers and Signal Processing, National Aerospace University named after N. Ye. Zhykovsky «KhAI», Kharkov, Ukraine, e-mail: nadejda_kozickaya@mail.ru.

Ponomarenko Nikolay Nikolaevich - Doctor of Technical Sciences, Associate Professor, Professor of Department of Transmitters, Receivers and Signal Processing, National Aerospace University named after N. Ye. Zhykovsky «KhAI», Kharkov, Ukraine, e-mail: nikolay@ponomarenko.info.