ORES
Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.75

МОДЕЛИРОВАНИЕ ТРАФИКА В СИСТЕМАХ ДОСТАВКИ КОНТЕНТА

Арсентьев Георгий Михайлович,
магистрант, Московский государственный технический университет им. Н.Э. Баумана (МГТУ им. Н.Э. Баумана)

Аннотация

В работе рассматриваются вероятностные модели трафика, порождаемого запросами пользователей к статическим файлам в сети Интернет. В отличие от традиционного анализа на уровне пакетов, внимание сосредоточено на двух ключевых характеристиках прикладного уровня: распределении вероятности запросов к файлам и распределении их размеров. Показано, что эмпирические данные не согласуются с классическими предположениями о равномерной популярности объектов и нормальном распределении их объёмов. Для моделирования популярности файлов предлагается использование закона Ципфа. Распределение размеров файлов описывается гибридной конструкцией, сочетающей логнормальное распределение для основной массы объектов и распределение Парето для больших файлов. Особое внимание уделено проблеме временной нестационарности трафика, проявляющейся в колебаниях интенсивности запросов и динамике популярности отдельных файлов. Интенсивность запросов предлагается представлять в виде произведения общего профиля активности пользователей и меняющейся во времени популярности отдельных файлов. Полученные результаты могут быть использованы при проектировании систем доставки контента, алгоритмов кеширования и генераторов синтетического трафика для имитационного моделирования.

Ключевые слова: моделирование трафика, статические файлы, распределение Ципфа, логнормальное распределение, распределение Парето, нестационарность, доставка контента.

TRAFFIC MODELING IN CONTENT DELIVERY SYSTEMS

Arsentyev Georgy Mikhailovich,
Master's Student, Bauman Moscow State Technical University (BMSTU)
e-mail: arsentevgm@student.bmstu.ru

ABSTRACT

This paper considers probabilistic models of traffic generated by user requests for static files on the Internet. In contrast to traditional packet-level analysis, the focus is placed on two key application-layer characteristics: the request probability distribution for files and their size distribution. It is shown that empirical data do not agree with the classical assumptions of uniform object popularity and normal size distribution. The Zipf law is proposed for modeling file

popularity. The file size distribution is described by a hybrid construct combining a lognormal distribution for the bulk of objects and a Pareto distribution for large files. Particular attention is paid to the temporal non-stationarity of traffic, which manifests itself in fluctuations of request intensity and the dynamics of individual file popularity. It is suggested to represent request intensity as the product of the overall user activity profile and the time-varying popularity of individual files. The obtained results can be applied in the design of content delivery systems, caching algorithms, and synthetic traffic generators for simulation modeling.

Keywords: traffic modeling, static files, Zipf distribution, lognormal distribution, Pareto distribution, non-stationarity, content delivery.

В условиях экспоненциального роста цифрового контента и стремительного развития сетевых технологий задача моделирования поведения пользователей на прикладном уровне приобретает первостепенное значение для проектирования и оптимизации современных информационных систем. Точные статистические модели востребованности контента и распределения его размеров необходимы для эффективного планирования пропускной способности каналов связи, построения интеллектуальных алгоритмов кеширования в сетях доставки контента (CDN), а также для реалистичного имитационного моделирования работы серверного оборудования [1].

Классические предположения о равномерной популярности файлов или нормальном распределении их размеров не соответствуют эмпирическим наблюдениям. Этот факт приводит к необходимости применения вероятностных моделей, более точно описывающих паттерны доступа к статическим файловым объектам в сети Интернет.

Центральной характеристикой поведения пользователей является распределение вероятности запроса файла. Если упорядочить все файлы по убыванию популярности и присвоить им ранги (самый популярный получает ранг 1, следующий получает ранг 2 и так далее), то эмпирические данные показывают, что вероятность запроса падает с ростом ранга по степенному закону, известному как закон Ципфа [2, с. 126–129]. В простейшей формулировке это означает, что вероятность запроса файла с рангом k обратно пропорциональна величине k^{α}, где показатель α — это положительное число, определяющее, насколько резко убывает популярность. Чем больше α, тем сильнее концентрация трафика на небольшой группе лидеров. Чтобы сумма вероятностей по всем N файлам равнялась единице, используется нормировочный делитель, равный сумме обратных степеней всех рангов от 1 до N.

Модифицированный вариант такого распределения может предполагать назначение рангов не отдельным файловым объектам, а их группам. Такой подход позволяет получить распределение для более глобальных процессов доступа к объектам (например, большие сети доставки контента или сеть Интернет).

Многочисленные исследования подтверждают, что частоты запросов к файлам подчиняются закону Ципфа или его модификациям. В одной из ранних работ на эту тему [2] отмечается, что практическим следствием такого распределения является существование крайне небольшого числа самых популярных файлов, генерирующих значительную часть всего трафика, и гигантского массива редко запрашиваемого контента.

Недавнее исследование глобального веб-трафика, охватывающее более 250 тысяч веб-ресурсов, показало, что частоты запросов подчиняются закону Ципфа с параметром формы $\alpha \approx 1.19$ [3]. Коэффициент Джини для такого распределения достигает 97%, что количественно подтверждает чрезвычайную концентрацию спроса. Эти результаты задают чёткие ориентиры для параметризации моделей трафика: при генерации синтетических

запросов необходимо воспроизводить именно такой степенной характер спада популярности.

На рисунке 1 показан график распределения вероятностей запроса к файлам по рангам в соответствии с законом Ципфа.

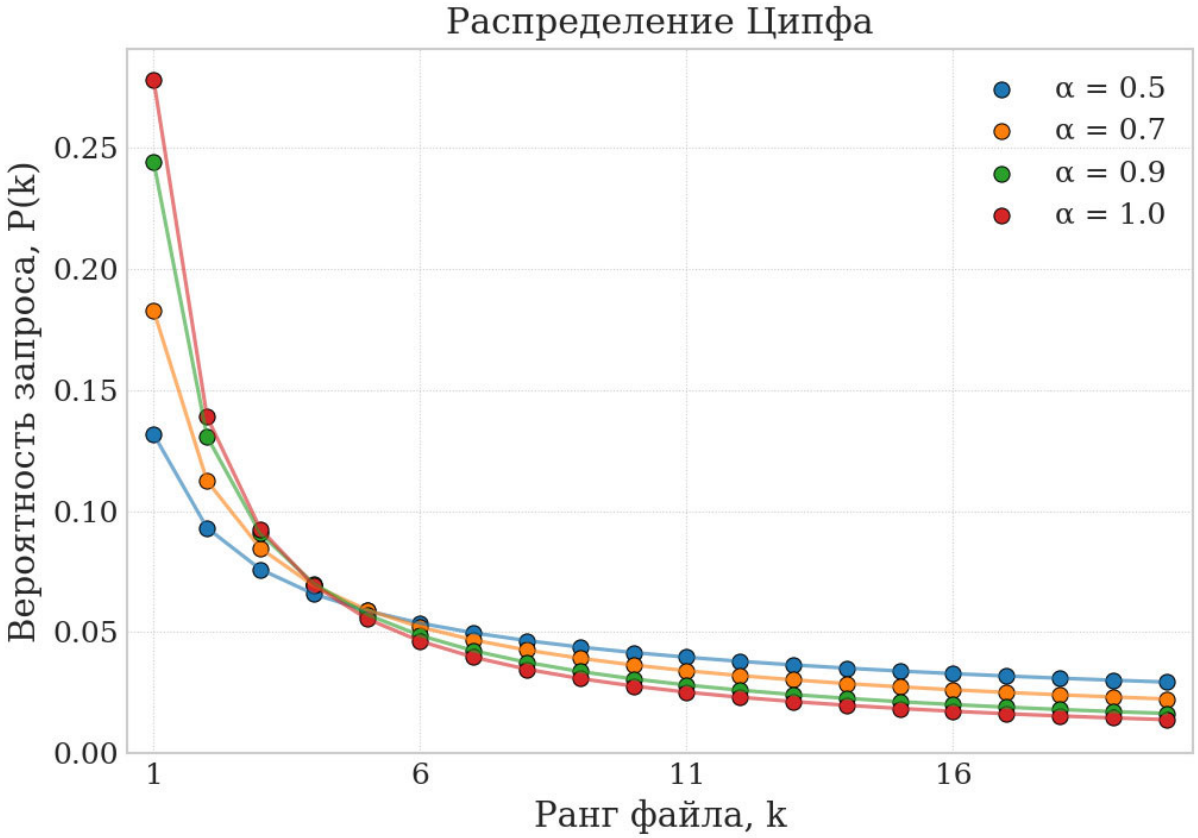

Рисунок 1 - Распределение вероятностей запросов к файлам [разработано автором]

Второй неотъемлемой характеристикой файла, критически важной для моделирования трафика, является его размер. Наблюдения показывают, что размеры статических файлов в Интернете распределены крайне неравномерно: в рамках одной пользовательской сессии могут запрашиваться как миниатюрные конфигурационные файлы размером в сотни байт, так и объёмные видеофайлы или архивы, достигающие нескольких гигабайт. Такая широкая вариативность исключает возможность использования простых симметричных распределений вроде нормального и требует привлечения моделей распределений с тяжелым хвостом.

Для описания основной массы файлов, имеющих относительно небольшие или умеренные размеры, хорошо подходит логнормальное распределение. Его ключевая особенность состоит в том, что нормальному закону подчиняется не сам размер, а его логарифм. Это приводит к асимметричной форме распределения с крутым подъёмом слева и более пологим спадом справа, что отражает мультипликативный характер процессов формирования объёма данных (например, последовательное редактирование, сжатие с переменным коэффициентом). Влиятельное исследование Доуни [4, с. 362–365], выполненное на обширном материале файловых систем и веб-серверов, показало, что логнормальная модель даёт хорошее приближение для подавляющего большинства файлов.

Однако логнормальное распределение неспособно адекватно описать «хвост» - область очень больших размеров, где сосредоточены редкие, но чрезвычайно объёмные объекты. Плотность логнормального закона в этой области убывает слишком быстро, делая

появление таких файлов практически невероятным с точки зрения модели, хотя в реальности они наблюдаются регулярно.

В современной научной литературе общепринятым стал гибридный подход, при котором распределение размеров файлов моделируется составной конструкцией: центральная часть описывается логнормальным законом, а правый хвост, начиная с некоторого порогового значения, - паретовским [5, с. 795–798]. Такая комбинация позволяет одновременно корректно воспроизводить как типичные объёмы передаваемых данных, так и редкие, но ресурсоёмкие всплески, обусловленные запросами к очень большим файлам.

На рисунке 2 показан график плотности вероятности распределения размеров файлов в соответствии с логнормальным законом.

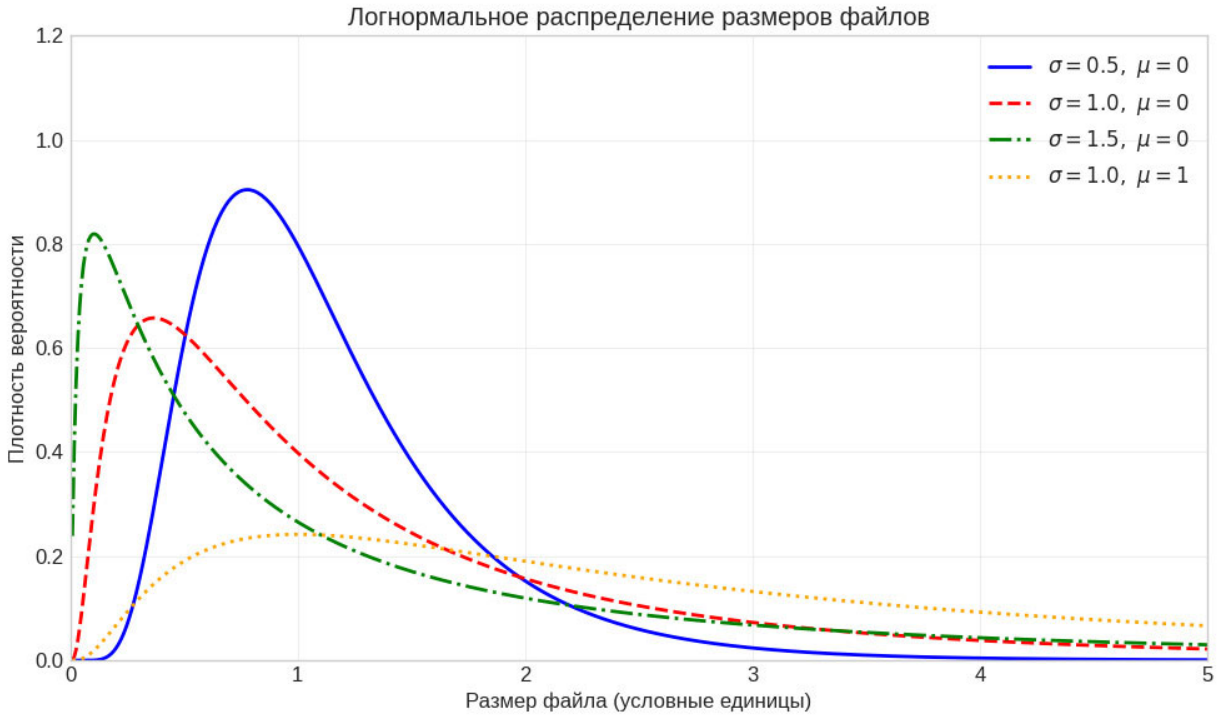

Рисунок 2 - Плотность вероятности распределения размеров файлов [разработано автором]

Ограничиться статическим описанием этих двух частных распределений было бы недостаточно, поскольку процесс запросов к файлам подвержен существенной временной нестационарности. Во-первых, в нём присутствуют детерминированные циклические компоненты: суточные и недельные колебания интенсивности обращений, обусловленные сменой активности пользователей [6]. Модели, не учитывающие такие циклы, будут систематически искажать оценки пиковых нагрузок и средней утилизации ресурсов. Во-вторых, популярность отдельных файлов не является постоянной. Она может резко возрастать (всплески интереса) и постепенно угасать. Даже на относительно коротких временных интервалах наблюдаются заметные флуктуации частот запросов, причём распределение этих флуктуаций стабильно для большинства объектов [3], что открывает возможность моделирования нестационарности через модуляцию параметров статического распределения. В-третьих, долгосрочные тренды, связанные с изменением состава контента и ростом числа пользователей, также вносят свой вклад в эволюцию статистических характеристик трафика.

Для учёта всей совокупности перечисленных нестационарных эффектов в практических задачах моделирования широко применяется подход, основанный на мультипликативной декомпозиции. Интенсивность потока запросов в произвольный момент времени представляется как произведение двух функций: глобального временного

профиля, который описывает изменение суммарной активности всех пользователей, и относительной функции популярности, которая задаёт распределение запросов между отдельными файлами в данный момент. Преимущество такого разделения состоит в том, что оно позволяет независимо настраивать циклическую и трендовую составляющие, не нарушая фундаментальных статистических свойств, таких как степенной характер закона Ципфа. На практике это означает, что параметр формы α распределения Ципфа может быть сделан функцией времени. Например, в часы наибольшей загрузки, α возрастает, отражая более резкую концентрацию запросов на лидерах, а в периоды спада уменьшается, делая распределение более равномерным.

Кроме того, для более тонкого воспроизведения изменчивости запросов к отдельным файлам в модель вводят случайные флуктуации вероятностей между соседними рангами. В реальных данных даже файлы с близкими рангами могут демонстрировать заметные расхождения в частоте обращений из-за кратковременных всплесков интереса, локальных трендов или различий в аудитории. Чтобы отразить этот эффект, в синтетический генератор запросов добавляют небольшой аддитивный или мультипликативный шум, который случайным образом увеличивает или уменьшает вероятность запроса для конкретного файла, не меняя общего рангового порядка. Такой приём позволяет избежать излишне жёсткой детерминированности модели и приблизить генерируемый трафик к наблюдаемому в реальных сетях, где популярность отдельных объектов никогда не следует идеально гладкой математической кривой.

Список литературы:

1. Fahrianto, F. Investigating Synthetic Traffic Generators for Zipf Distribution Simulation Accuracy / F. Fahrianto, H. B. Suseno, H. T. Ciptaningtyas [et al.] // INFOTEL. – 2025. – Vol. 17, no. 2. – P. 268–278.
2. Breslau, L. Web caching and Zipf-like distributions: evidence and implications / L. Breslau, P. Cao, L. Fan [et al.] // Proc. of the Eighteenth Annual Joint Conf. of the IEEE Computer and Communications Societies (INFOCOM ’99). – New York, 1999. – Vol. 1. – P. 126–134.
3. Xavier, H. S. The Web unpacked: a quantitative analysis of global Web usage / H. S. Xavier // arXiv preprint arXiv:2404.17095. – 2024. – DOI: 10.48550/arXiv.2404.17095.
4. Downey, A. B. The structural cause of file size distributions / A. B. Downey // Proc. of the 2001 ACM SIGMETRICS Intern. Conf. on Measurement and Modeling of Computer Systems. – Cambridge, MA, USA, 2001. – P. 361–370. – DOI: 10.1145/378420.378824.
5. Gong, W. Lognormal and Pareto distributions in the Internet / W. Gong, Y. Liu, V. Misra, D. Towsley // Computer Communications. – 2005. – Vol. 28, no. 7. – P. 790–801. – DOI: 10.1016/j.comcom.2004.11.001.
6. Kirci, E. C. Five Blind Men and the Internet: Towards an Understanding of Internet Traffic / E. C. Kirci, A. Mishra, L. Vanbever // 1st New Ideas in Networked Systems (NINeS 2026). – OASIcs. – 2026. – Vol. 139. – Article 25. – P. 25:1–25:26.

References:

1. F. Fahrianto, H. B. Suseno, H. T. Ciptaningtyas, et al., "Investigating synthetic traffic generators for Zipf distribution simulation accuracy," INFOTEL, vol. 17, no. 2, pp. 268–278, 2025.

2. L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in Proc. IEEE INFOCOM '99, New York, NY, USA, 1999, vol. 1, pp. 126–134. doi: 10.1109/INFCOM.1999.749260.
3. H. S. Xavier, "The Web unpacked: A quantitative analysis of global Web usage," arXiv preprint arXiv:2404.17095, 2024. doi: 10.48550/arXiv.2404.17095.
4. A. B. Downey, "The structural cause of file size distributions," in Proc. ACM SIGMETRICS 2001, Cambridge, MA, USA, 2001, pp. 361–370. doi: 10.1145/378420.378824.
5. W. Gong, Y. Liu, V. Misra, and D. Towsley, "Lognormal and Pareto distributions in the Internet," Comput. Commun., vol. 28, no. 7, pp. 790–801, 2005. doi: 10.1016/j.comcom.2004.11.001.
6. E. C. Kirci, A. Mishra, and L. Vanbever, "Five blind men and the Internet: Towards an understanding of Internet traffic," in Proc. 1st New Ideas in Networked Systems (NINeS 2026), OASIcs, vol. 139, art. 25, pp. 25:1–25:26, 2026.