Исследование пользовательских предпочтений для контроля и оптимизации Интернет-трафика в организации

© Леонова Ю.В., Федотов А.М.

Институт вычислительных технологий СО РАН
juli@ict.nsc.ru

Аннотация

В статье рассмотрены вопросы, связанные с оптимизацией потребления интернет-трафика и повышения эффективности работы интернет-канала, описаны проблемы управления информационными ресурсами и их защиты, основные принципы и понятия, методика поэтапного сокращения затрат на Интернет-трафик и потерь рабочего времени, связанных с нецелевым использованием сети Интернет, на примере научно-образовательной сети ННЦ СО РАН.

Введение

В крупных городах многие провайдеры уже предоставляют возможность безлимитного использования интернет-канала за фиксированную сумму, однако стараются ограничивать либо пропускную способность предоставляемого канала, либо требуют соблюдения некоторых условий, направленных на то, чтобы удержать потребление клиентом интернет-трафика в определенных рамках. Кроме того, цены на безлимитный доступ в большинстве случаев все еще "кусаются".

Помимо стоимости услуг интернет-провайдеров, отдельно стоит вопрос состава потребляемого трафика. Редкий руководитель, подписывая очередной счет, не задавался вопросом: а за что, собственно, платятся деньги? Действительно ли оправданы затраты и есть ли способ их снижения?

Ниже мы рассмотрим вопросы оптимизации потребления интернет-трафика и повышения эффективности работы интернет-канала.

1 Проблемы управления информационными ресурсами и их защиты

Одной из особенностей Интернета является то, что на определенном этапе он развивался стихийно. Это, с одной стороны, обеспечило массовый характер его использования, а с другой — породило ряд проблем с серьезными последствиями.

1) Поскольку Интернет является каналом во внешний мир, он стал основным источником распространения вредоносного мобильного кода (вирусов, червей, троянских программ).

2) Глобальная сеть стала использоваться в качестве канала, через который осуществляются атаки на локальные вычислительные сети организаций, отдельные серверы и компьютеры. Многие Интернет-ресурсы включают в себя различный программный код — JavaScript, Flash, ActiveX и другие. Злоумышленники могут эксплуатировать этот код для организации атак на корпоративные сети и пользовательские рабочие места.

3) В настоящее время Интернет может рассматриваться как один из основных каналов утечки конфиденциальной информации. Например, информационные ресурсы компаний подвергаются серьезным угрозам из-за использования сотрудниками этих компаний бесплатных почтовых ящиков. Многие сотрудники различных компаний помимо внутренних корпоративных почтовых адресов активно используют бесплатные почтовые ящики, предоставляемые различными провайдерами. Имея доступ к Интернету со своего рабочего места и зная, что канал не контролируется, любой пользователь может беспрепятственно отправить за пределы организации любую конфиденциальную информацию. Но, даже понимая это, не каждая компания запрещает использовать бесплатные почтовые сервисы, тем самым, позволяя своим сотрудникам решать, как и какую информацию отправить за пределы компании.

4) Бесконтрольный доступ к Интернету значительно снижает производительность труда в коллективе. Простота освоения, легкость поиска необходимой информации и другие полезные качества Интернета — вот причины того, что данный сервис широко применяется, в том числе и для личных целей. Не секрет, что у многих уже давно появилась привычка начинать рабочий день с чтения новостей, просмотра сводок погоды и т.п.

Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.

Сотрудники различных организаций и компаний используют Интернет в целях, не имеющих прямого отношения к их работе. Это и "походы" в Интернет-магазины, и сетевые игры, и просто поиск информации.

5) Наконец, еще одно следствие неконтролируемого использования Интернета — это снижение пропускной способности сети. Сотрудники организаций используют корпоративные ресурсы для просмотра видео, прослушивания аудиозаписей (через потоковые аудио- и видеоканалы), играют в сетевые игры, загружают файлы большого объема (например, файлы мультимедиа: графические, музыкальные файлы, фильмы и т.п.), что создает значительную нагрузку на локальные вычислительные сети.

Таким образом, проблемы управления информационными ресурсами вычислительных сетей и их защиты становятся все более актуальными для организаций.

2 Контроль и оптимизация Интернет-трафика

2.1 Аудит сети

Первым шагом для решения вышеперечисленных проблем является аудит сети организации, что позволит выявить "дыры" и узкие места в компьютерной системе организации, в том числе даст картину потребления интернет-трафика. В результате аудита можно получить не только данные о том, что происходит в сети организации, в каком состоянии находятся ее ресурсы, какова структура интернет-трафика, но и о том, чем конкретно занимаются сотрудники на рабочих местах.

2.2 Контроль доступа

Следующим шагом по сокращению затрат на интернет-трафик может стать контроль доступа к Интернет-ресурсам, который можно решить двумя способами. 1) Запрещение использования Интернета без необходимости, когда пользователям разрешается доступ только к строго определенным сайтам. 2) Контроль действий сотрудников, при этом сотрудник может свободно пользоваться ресурсами Интернета. Но если пользователь выполнит действия, противоречащие политике безопасности, это будет обнаружено и пресечено. Второй способ контроля является наиболее гибким и более распространенным, но именно при его применении возникают существенные проблемы, которые состоят в том, что практически невозможно однозначно определить, к какой информации следует запретить доступ. Необходимой составляющей решения этих проблем является разработка и внедрение политики безопасности сети и политики использования ресурсов.

2.2.1 Политики безопасности

Политика безопасности сети – это набор законов, правил и практических рекомендаций, на основе которых строится управление, защита и распределение защищаемых информационных ресурсов. Она должна охватывать все особенности процесса использования информационных ресурсов сети организации, определяя поведение системы в различных ситуациях. Ключевым шагом разработки политики безопасности является определение критичных для организации ресурсов и возможных угроз доступности, конфиденциальности и целостности этих ресурсов. При этом может применяться несколько подходов, в том числе ранжирование сетевых ресурсов по их стоимости, по вероятности реализации угроз и по серьезности их последствий для организации. Последняя не всегда связана с раскрытием конфиденциальной информации или выходом из строя дорогостоящих устройств, но также снижением производительности ресурсов, которые активно используются сотрудниками организации при исполнении служебных обязанностей. Поэтому выявление таких сетевых информационных ресурсов является важной задачей при разработке политики безопасности.

Существует несколько решений этой задачи, но наиболее эффективным представляется прослеживание истории сетевых взаимодействий путем накопления и анализа статистики обращений к сетевым серверам и предоставляемым ими сервисам. Сохраняя в базе данных структурированную информацию по сетевому трафику, извлеченную из заголовков передаваемых по сети пакетов, и обрабатывая накопленные данные с помощью автоматизированных алгоритмов анализа, можно составить четкую схему использования информационных ресурсов сети, необходимую для формирования сбалансированной политики безопасности сети, без выполнения трудоемкой и рутинной ручной работы. При этом рассмотренный механизм позволяет не только учесть сетевые взаимодействия внутри организации, но и обращения к информационным ресурсам из внешних по отношению к организации источников, в частности удаленный доступ к ресурсам сети, и таким образом выявить источники потенциальных угроз.

2.2.2 Политика использования Интернет-ресурсов

Для обеспечения гибкости контроля использования Интернет-ресурсов в организации вводится политика использования ресурсов. Эта политика может реализовываться на основе анализа и фильтрации веб-трафика. На сегодняшний день существует множество как коммерческих, так и некоммерческих решений. К наиболее распространённым коммерческим продуктам можно отнести: open-source систему Poesia [4],

коммерческие системы CyberPatrol [2], SurfControl [5], NetNanny [3] и др.

Можно выделить два основных признака систем фильтрации и анализа трафика – способ и время анализа трафика. По способу анализа все системы можно разбить на два больших класса: 1) анализирующие лишь общую (мета-) информацию о ресурсе; 2) анализирующие в том числе и содержимое (контент) ресурса.

По времени анализа все системы можно также разбить на два класса: 1) анализирующие информацию в реальном времени (онлайн), т.е. во время запроса пользователем Интернет-ресурса; 2) анализирующие информацию в отложенном режиме (оффлайн), т.е. после того, как пользователь получил доступ к ресурсу.

В данной работе рассматриваются системы масштаба локальных сетей, анализирующие метаинформацию в отложенном режиме.

Применение системы контроля использования Интернет-ресурсов нельзя представить без анализа событий, происходящих в системе. Администраторам необходимо оперативно получать информацию о текущем состоянии системы, а также сводные отчеты об использовании Интернет-ресурсов пользователями или группами пользователей. Такая информация позволяет не только контролировать использование Интернет-ресурсов, но и проверять эффективность политики безопасности и динамически адаптировать ее к изменяющимся условиям и задачам. Поэтому в большинстве существующих средств контроля использования Интернет-ресурсов есть возможность формирования статистических отчетов, а также интерактивного наблюдения за доступом к внешним ресурсам.

Существует несколько способов представления статистических данных о трафике. Первый способ предполагает использование внутренних возможностей продукта, то есть встроенной системы генерации отчетов. Как правило, в состав такой системы входят подсистема генерации отчетов и база данных, в которую в виде журналов записывается вся информация о событиях, а также некоторые запросы пользователей (например, команды POST). С помощью внутренних средств производятся SQL-запросы к базе данных, результаты которых дают наглядную картину трафика и действий пользователей. При этом могут создаваться типовые запросы (например, "100 часто загружаемых сайтов", "100 пользователей, переславших наибольшие объемы данных за указанный период", "100 самых активных пользователей" и т.п.) с изменяемыми параметрами (например, по дате и времени). Другой способ предполагает получение отчетов с помощью стандартных средств, таких как Crystal Reports, Oracle Reports и т.п. Эти средства интегрируются с системой контроля использования Интернет-ресурсов и тоже используют базу данных, которая создается в результате фильтрации трафика.

2.3 Установка кэширующих серверов и зеркал

Еще одним способом оптимизации интернет-трафика является использование кэширующих серверов и системы зеркал, на которые в «прозрачном» для конечного пользователя режиме перенаправляются HTTP-запросы пользователей. Использование кэширующих серверов и системы зеркал преследует две основные цели:

- Улучшение производительности: снижение нагрузки на каналы провайдера, используемые для выхода в интернет и уменьшение времени ожидания загрузки данных для пользователей.
- Сокращение затрат: размер трафика на канал в интернет после установки кэширующих серверов и зеркал уменьшится, что приведет к снижению платежей за передачу информации по этим каналам.

1) Создание системы кэширования интернет-трафика позволяет увеличить пропускную способность канала связи, одновременно снизив среднее время ожидания ответа на запрос пользователя. Кэширование минимизирует задержки при передаче файла, примерно в 5-200 раз. Суть кэширования www-трафика состоит в том, что запрос пользователя на получение документа перенаправляется на кэш-сервер, который сначала проверяет наличие документа в своем кэше, после чего продолжает обслуживание запроса. Если документ в кэше не найден, то кэш-сервер направляет запрос на сервер-источник документа или другому кэш-серверу.

Система кэширования с иерархической сетью создана, например, в Научном центре в Черноголовке, объединенной с кэш-серверами в Ярославле, Перми, Челябинске, МНФ (Москва), ИТФ им. Л.Д.Ландау (Москва), ИОХ им. Зелинского (Москва) [8].

2) Развертывание зеркал позволяет разместить наиболее востребованные данные «ближе» к пользователю. Зеркала обеспечивают максимальную скорость передачи данных от сервера к пользователю – при запросе файла с веб-сайта, его передает локальное зеркало. Под зеркалированием интернет ресурсов понимается создание полных или частичных копий (зеркал) этих ресурсов на географически удаленных серверах, обновление которых может производиться во время минимальной загрузки каналов, например, ночью. Прозрачное перенаправление на зеркало (незаметное для пользователя) реализуется посредством редиректора, который выполняет первоначальную обработку URL, и либо возвращает прежний URL для дальнейшей обработки прокси-серверу в случае, если все в порядке, либо возвращает тот, который, по его мнению, более правильный.

Естественно, что сам процесс зеркалирования создает определенную нагрузку на центральный

сервер и каналы связи (порой сравнимую, а иногда и превышающую выигрыш от зеркал). Зеркалирование увеличивает общую сложность системы (проблемы с администрированием, распределением прав, увеличением технического парка и т.д.). Можно утверждать, что необдуманное внедрение зеркал приведет к негативному результату. С другой стороны зеркала могут ощутимо повысить надежность и общую производительность системы.

Система зеркал была реализована при создании региональной научно-образовательной сети в Интернет центре Новгородского государственного университета [6], что позволило уменьшить внешний трафик организации.

В рамках данной работы было проведено исследование статистики обращений к веб-серверам, на основе которого рассмотрены вопросы, связанные с оптимизацией трафика и ускорением работы Интернет, в том числе задание и приложение правил обслуживания и учета трафика HTTP прокси-сервером, а также задание и реализация политики безопасности. Выработка и внедрение корпоративной сетевой политики, основной принцип которой сводится к тому, что пользователи научно-образовательной сети работают в первую очередь с научно-образовательной информацией.

3 Исследование Web-трафика

С момента своего появления технология веб стала предметом исследований [1]. Основная цель большинства исследований в вебе – это поиск таких свойств трафика, которые позволят совершенствовать саму технологию, увеличить скорость передачи информации к пользователю, уменьшить время загрузки нужного документа. Повышенный интерес к исследованиям веб-трафика вызван тем, что в настоящее время веб-трафик доминирует в общем трафике всех компьютерных сетей.

В одной из первых работ по исследованию веб-трафика [7] было замечено, что популярность документов в вебе распределена очень неоднородно. Большинство запросов приходят на очень небольшое количество документов, в то время как многие документы запрашиваются всего несколько раз. Для описания свойств популярности веб-документов очень удобно использовать технику ранговых распределений.

Рассмотрим информацию, которую можно получить на основе анализа логов прокси-сервера.

1. Информационный ресурс. Информационный ресурс представляет собой совокупность информационных объектов. Основные параметры информационного объекта: тип информации — текст, изображения, аудио-, видеоданные, потоковые данные, бинарные файлы, медиаданные; объем информации; в) приемлемая скорость доступа к объекту; полезность; частота модификации; потребность в объекте; права доступа на объект.
2. Потребитель ресурса — пользователи или
3. компьютеры. Основные параметры потребителя ресурса: текущее и потенциальное количество пользователей ресурса; интенсивность запросов к каждому информационному объекту, объем потребления, генерируемый трафик; удовлетворенность качеством доступа.
3. Канал передачи данных. Канал передачи данных между информационным ресурсом и потребителем. Основные параметры: полоса пропускания; загрузка канала (входящий/исходящий трафик); доля трафика ресурса в общей загрузке канала; стоимость работы по каналу.

Анализ полученной информации может быть использован для решения следующих задач.

1) **Оптимизация (уменьшение) трафика.** Как правило, наиболее «узким» местом является внешний канал научно-образовательной сети, когда большое количество пользователей одновременно работает с разнообразными Интернет-ресурсами и возникает перегрузка канала. Решение этой проблемы заключается в кэшировании и классификации наиболее важных и востребованных информационных ресурсов, например статей, с последующим размещением их для использования научным сообществом внутри локальной сети. В результате при уменьшении количества перекачек повышается надежность сети.

2) **Изучение информационных потребностей.** Данный анализ позволяет получить информацию о поведении пользователей локальной сети в Интернете, выявлять самых активных пользователей и смотреть, какие ресурсы они посещают, получать общее представление о распределении трафика по сайтам, дням недели и времени суток и многое другое. При обнаружении наиболее напряженных участков скачивания «важных» ресурсов может быть увеличена пропускная способность на данном направлении.

3) **Ограничение нецелевого использования.** Большой эффект по разгрузке канала дает ограничение трафика с нежелательным содержанием, например, порно-сайтов или развлекательных ресурсов типа «Одноклассники», различных «непрофильных» ресурсов аудио- и видео-серверов.

Установка или настройка существующих корпоративных прокси-серверов позволяет уменьшить внешний трафик организации и повышает качество работы с ресурсами. Для этого производится дополнительная настройка прокси-серверов: ограничивают доступ к непрофильным серверам; вводят ряд ограничений по пропуску типов файлов (avi, mp3 и т. д.); ограничивают пользователей по скорости доступа; при необходимости увеличивают размер кэша; при необходимости изменяют время хранения документов в кэше востребованных ресурсов.

Обозначение	Кэш-сервер	Период	Число запросов	Число сайтов	Объем
ICT1	proxy.ict.nsc.ru	3 недели	20,250,832	104,249	337,29 G
ICT2	proxy.ict.nsc.ru	2 недели	11,402,797	63,190	240.09 G
NSU1	proxy.nsu.ru	2 недели	34,040,909	113,324	276,52 G
NSU2	proxy.nsu.ru	2 недели	32,908,553	121,999	253,42 G

Таблица 1. Набор данных

3.1 Исследование наборов данных

Напомним основной принцип работы протокола HTTP. Для того, чтобы получить нужный документ, пользователь направляет запрос к веб-серверу, на котором находится этот документ. Веб-сервер в ответ возвращает пользователю требуемый документ. Кроме того, пользователь может посылать запрос не напрямую к веб-серверу, а на сервер-посредник, с которым у него имеется высокая скорость соединения (например, к прокси-серверу в его локальной сети). Веб прокси-сервер, как правило, имеет кэш и, если запрашиваемый документ находится в кэше прокси-сервера, то скорость получения этого документа значительно возрастает.

Таким образом, для исследователей имеется несколько способов получения информации о веб-трафике. Можно исследовать запросы, приходящие к отдельно взятым веб-серверам, можно собирать информацию о действиях отдельных пользователей или анализировать запросы пользователей к кэш-серверам. Основное отличие между этими способами состоит в том, что в первом случае мы получаем данные о трафике для очень небольшого подмножества веб, которым является множество документов на нескольких выбранных веб-серверах, а информация, полученная из машин пользователей или прокси-серверов, дает нам представление о трафике, создаваемым небольшой группой пользователей.

Поскольку нас интересовали исследование свойств для запросов второго вида, в качестве источника информации для исследования веб-трафика были взяты логи информация кеш-сервров. Для анализа использовались лог-файлы кэш-серверов сети ННЦ СО РАН: proxy.ict.nsc.ru (СО РАН) и proxy.nsu.ru (НГУ) – типичные прокси-серверы, обслуживающие запросы локальных пользователей организации. Мы проанализировали данные, собранные в течение одного месяца. Детальное описание наборов данных дано в таблице 1.

На рассматриваемых серверах установлено программное обеспечение Squid. Его лог представляет собой текстовый файл, в который записывается информация о всех запросах, поступивших на кэш-сервер. После получения очередного запроса кэш-сервер добавляет в лог-файл одну строку с информацией, характеризующей полученный запрос, например

1210274283.328 1010 194.226.177.55 TCP_HIT/200 67526 GET http://stats.iihf.com/Hydra/132/IHM132000_85K_6_0.pdf - DIRECT/80.231.19.71 application/pdf

Здесь 1210274283.328 обозначает время поступления запроса в формате UTC (Universal coordinated time), 1010 обозначает, сколько времени (в мс) заняла обработка запроса, 194.226.177.55 – IP адрес машины пользователя, пославшего запрос, TCP_HIT/200 – код результата выполнения запроса, 599 – размер запрашиваемого ресурса (в байтах), GET – метод протокола HTTP. Большинство запросов к кэш_серверу используют метод GET – метод получения нужного ресурса по протоколу HTTP.

http://stats.iihf.com/Hydra/132/IHM132000_85K_6_0.pdf – адрес запрашиваемого ресурса,

application/pdf – MIME-тип документа, в данном случае документ в формате PDF.

Далеко не все запросы пользователей к кэш-серверу благополучно им обрабатываются, например кэш-сервер может быть настроен таким образом, что он обрабатывает запросы только от определенной группы пользователей, а остальные запросы игнорирует. В другом случае, пользователь может сделать ошибку при вводе URL, запросить несуществующий документ или документ, для получения которого необходимо ввести пароль, который пользователь вводит неверно. Наконец, во время передачи данных может просто разорваться связь. Результат обработки запроса (код HTTP) кэш-сервер заносит в соответствующее поле лог-файла. Для того, чтобы достоверно судить о скачиваемости документов, мы будем анализировать только запросы, успешно обработанные кэш-сервером, имеющие код результата выполнения – 200. Таким образом, на втором этапе обработки данных мы оставляем только те записи в лог-файлах, у которых в поле результата выполнения запроса записано 200. На следующем этапе обработки данных мы выделяем из полей URL документа. Затем мы подсчитываем количество появлений каждого документа в лог-файле – f_i и, сортируя документы по убывающим значениям f_i, мы получаем ранговое распределение популярности скачивания документов. Выделяя из поля URL название веб-сайта, аналогично можно определить популярность веб-сайтов. Аналогично определяется ранговое распределение объема скачивания документов. Далее для определения предпочтений пользователей выполняется категоризация полученных данных по областям деятельности в два этапа. На первом этапе категоризикация выполняется на основе классификатора каталога сайтов Яндекса.

ПК, Интернет, связь	Здоровье	Театры
Hardware	Мода и красота	Фотография
Интернет	Покупки	Музеи
Мобильная связь	Общество	Изобразительные искусства
Программы	Власть	Танец
Безопасность	Законы	Спорт
Сети и связь	НКО	СМИ
Интерфейс	Политика	Периодика
Работа	Религия	Информационные агентства
Учеба	Развлечения	Телевидение
Высшее образование	Игры	Радио
Курсы	Юмор	Бизнес
Среднее образование	Непознанное	Финансы
Школы	Личная жизнь	Недвижимость
Науки	Отдых	Строительство
Учебные материалы	Где развлечься	Производство и поставки
Дом	Туризм	Реклама
Квартира и дача	Хобби	Деловые услуги
Кулинария	Культура	Все для офиса
Все для праздника	Музыка	Справки
Семья	Литература	Транспорт
Домашние животные	Кино	Афиша
		Авто

Таблица 2. Классификатор Яндекса

Каталог Яндекса, содержащий описания сайтов русскоязычного интернета, систематизированных по тематическим категориям, построен на основе фасетной классификации. Такая классификация с одной стороны позволяет легко организовать поиск ресурсов не только по тематике, но и по типу информации, а с другой стороны предотвращает углубление рубрикатора и неоднозначность тематического отнесения ресурсов. На первом уровне дерева каталога имеется 13 тем, а число уровней в глубину не превышает четырех. Рубрики сгруппированы определенным образом. В первой группе темы «человек и его окружение»: дом, учеба, работа, общество, коммуникации. Вторая группа – «развлечения»: отдых, юмор, спорт, музыка и др.

Третья группа – «бизнес и экономика». Зато, помимо тем, в каталоге имеется ряд дополнительных признаков (фасет), позволяющих уточнить характер ресурсов, которые пользователь хочет увидеть в тематических категориях. Эти нетематические признаки характеризуют ресурсы по региону, сектору экономики, степени достоверности (источнику) информации, ее потенциальной аудитории (адресату информации), жанру (художественная литература, научно-техническая литература, и т. д.), цели (предложение товаров и услуг, интернет-представительство) и т. д.

Выше приведен классификатор Яндекса (табл.2).

Посредством программы-робота, запрограммированного на обход каталога сайтов Яндекса по различной глубине вложенности был сформирован классификатор, на основе которого выполнялась категоризация трафика по областям деятельности. На втором этапе выполняется категоризация оставшегося трафика на основе сигнатурного подхода, основанный на использовании экспертной базы знаний адресов Интернет-ресурсов. Такая база знаний содержит адреса ресурсов, с каждым из которых связан набор тем (категорий), к которым, по мнению экспертов, относится данный Интернет-ресурс. Для категоризации трафика был разработан классификатор доменных имен, с рубрикацией, аналогичной классификатору Яндекса. Полученные результаты исследования приведены на рис.1.

3.2 Обработка результатов

Можно заметить, что ранговое распределение предпочтений пользователей для различных кэш-серверов различаются, но близки друг к другу. Массовыми категориями, на долю которых приходится основной объем трафика являются для ИВТ СО РАН “ПК, Интернет, связь”, “Новости, СМИ“ – 58,61%, а для НГУ “ПК, Интернет, связь”, “Культура”, “Развлечения”, “Справки” – 74,15%.

Ясно, что для самых массовых категорий: “Культура”, “Развлечения” кэширование не является оправданным, поскольку наиболее популярными являются сервисы предоставляющие мультимедийные услуги, такие как просмотр флэш-, видеороликов, прослушивание радио и музыкальных файлов, использование других клиентских приложений, которые в своей работе используют передачу динамической информации. Категория “Новости, СМИ” также содержит в большей степени информацию динамического характера. В этом случае доля статической (кэшируемой) информации составляет малую часть общего веб-трафика. И эффективность использования кэширования – сводится к нескольким единицам процентов.

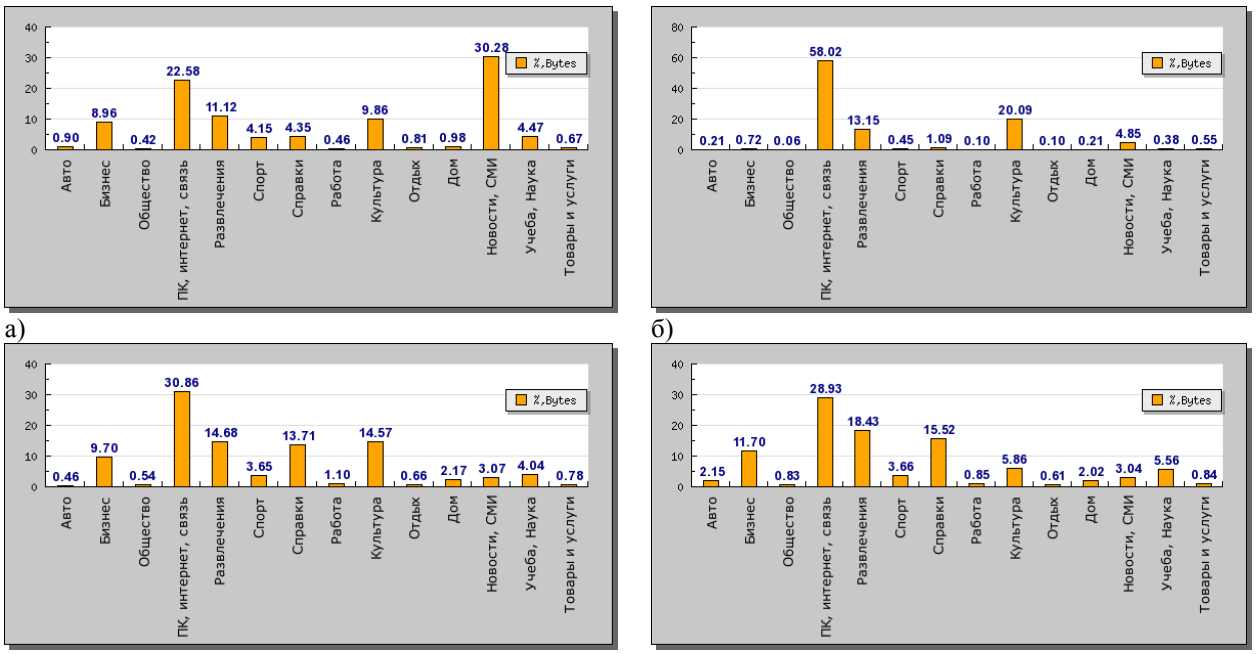


Рис.1 Ранговое распределение объема трафика по категориям для кэш-серверов
а) ICT1 б) ICT2 в) NSU1 г) NSU2

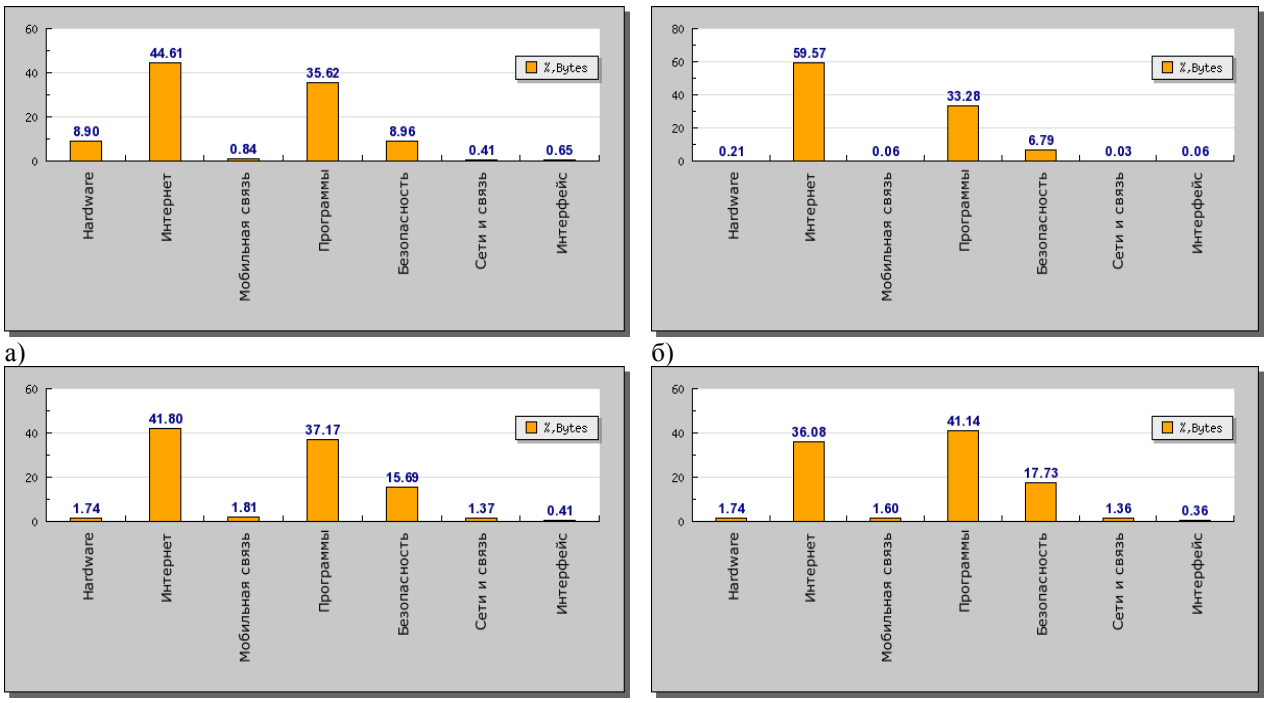
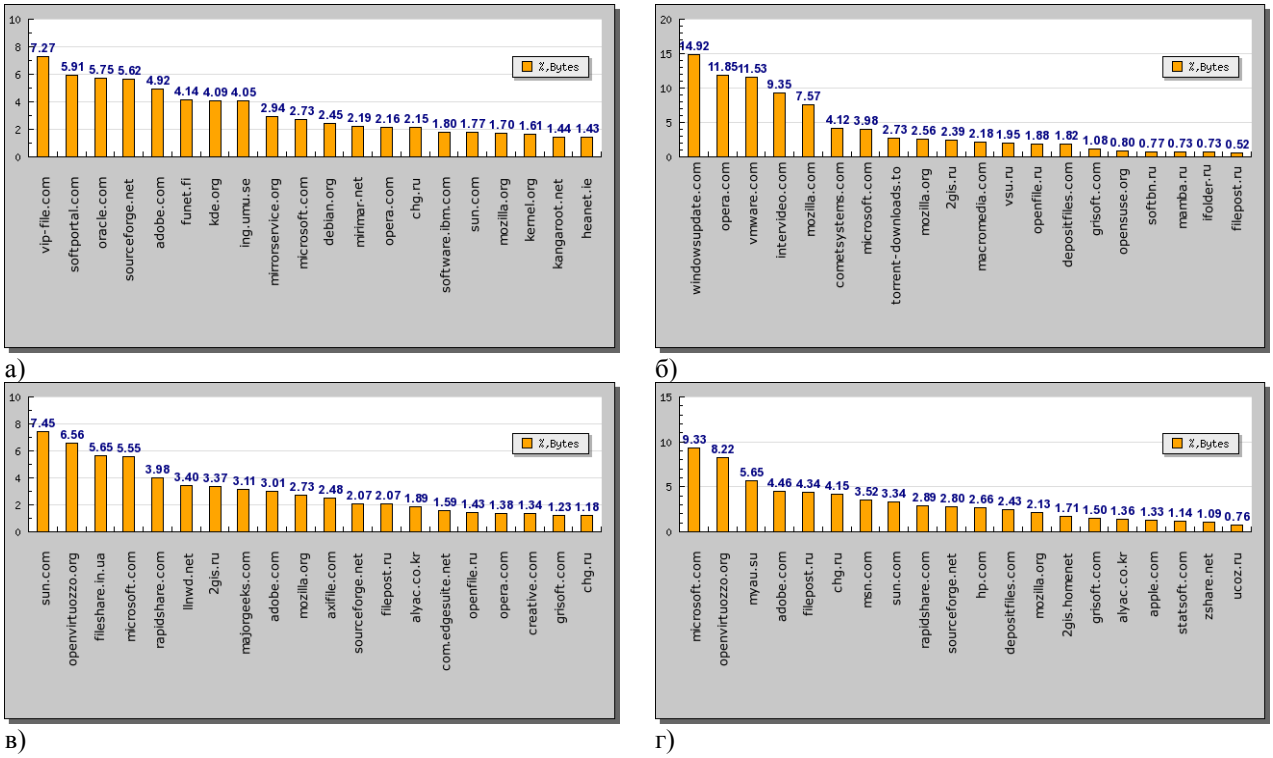


Рис.2 Ранговое распределение трафика в категории “ПК, Интернет, связь”
а) ICT1 б) ICT2 в) NSU1 г) NSU2

Очевидно, что существенной экономии трафика можно добиться кэшированием ресурсов статического характера, как архивы, полные тексты, программное обеспечение и т.п. Поэтому для дальнейшего анализа была выбрана наиболее массовая категория “ПК, Интернет, связь”, на долю которой приходится более 30% трафика. Проведем дальнейшую детализацию данной категории (рис.2).

Видно, что подкатегориями, на которые приходится основной трафик, являются “Интернет”, “Программы” и “Безопасность”. В подкатегории “Интернет” преобладают динамические ресурсы – электронная почта, запросы к поисковым системам, баннерные сети, счетчики и рейтинги. В подкатегории “Программы” преобладают статические ресурсы – программное обеспечение, а



а) б)

в) г)

Рис.3 Ранжирование сайтов по объему трафика (20 хитов)
а) ICT1 б) ICT2 в) NSU1 г) NSU2

подкатегория “Безопасность” также содержит статические ресурсы – антивирусное ПО, ПО для защиты от взлома и спама и т.п.

Дальнейший анализ выявил наиболее «объемное» скачиваемое ПО и его обновления:

- 2GIS
- Adobe Acrobat
- Adobe Macromedia flash
- Adobe photoshop
- CentOS
- Debian
- Eclipse
- FCKeditor
- JRE, JDK
- KDE
- Linux fedora
- Lotus
- Miktex
- Mozilla
- Nero
- Opera
- Oracle
- Pascal
- Safari
- Suse
- Tinymce
- VMWare
- Windows XP, Vista

На основе данного списка ресурсов будет приниматься решение о целесообразности создания зеркала для конкретного информационного ресурса, исходя из стоимости создания зеркала и поддержки его функционирования, возможностей зеркалирования ресурса, организационных, административных и юридических аспектов. В настоящее время в ИВТ СО РАН создан зеркальный сервер, содержащий обновления Windows XP, на который осуществляется «прозрачное» перенаправление пользователей.

Далее было произведено ранжирование сайтов подкатегории “Программы” по объему трафика (рис.3).

4 Заключение

Перечисленные методы сокращения затрат на интернет-трафик и потерь рабочего времени, связанных с нецелевым использованием сети интернет, могут быть использованы в любой компании. Эффект от их использования может быть различным. В одном случае затраты могут сократиться на 10%, в другом - в 2-3 раза.

Авторы благодарят рецензентов, высказанные замечания будут учтены в докладе на конференции.

Литература

[1] A caching Relay for the World Wide Web. Proc. 1st International Conference on the World Wide Web, CERN, Geneva (Switzerland), May 1994. Elsevier Science, p. 69–76.
[2] CyberPatrol Internet Security Software. http://www.cyberpatrol.com/
[3] NetNanny Parental Control. http://www.netnanny.com/
[4] Open-Source Filtering Software. http://www.poesia-filter.org/
[5] SurfControl url and keyword-based Internet filtering and blocking software. http://www.surfcontrol.com/

[6] Герасимов В.В., Курмышев Н.В. Типовой проект создания регионального зеркала. // В сборнике научных статей "Интернет-порталы: содержание и технологии". Выпуск 3. / Редкол.: А.Н. Тихонов (пред.) и др.; ФГУ ГНИИ ИТТ "Информика". - М.: Просвещение, 2005. - С. 379-392.

[7] Крашаков С.А., Теслюк А.Б., Щур Л.Н. Об универсальности рангового распределения популярности веб-серверов // Вестник РФФИ – 2004 - № 1 - с. 46-66.

[8] Крашаков С.А., Щур Л.Н.. Кеширование информационных потоков и стратегия оптимизации маршрутов в распределенных системах. Тезисы докл. 2-ой Всерос. конф. “Научный сервис в сети Интернет”, Новороссийск, сент. 2000, с. 145-148.

Research of the user preferences for the control and Internet traffic optimisation in the organisation

Leonova Yu., Fedotov A.

In article the questions connected with optimisation of consumption of the Internet traffic and increase of an overall performance of the Internet channel are considered, management problems by information resources and their protection, main principles and concepts, a technique of stage-by-stage reduction of expenses for the Internet traffic and the losses of working hours connected with no-purpose use of a network the Internet, on an example of a scientifically-educational network of NSC SB RAS are described.